Jens Clausen
Neil Levy
Editors

Handbook of Neuroethics

Springer Reference

Handbook of Neuroethics

Jens Clausen • Neil Levy
Editors

Handbook of Neuroethics

With 31 Figures and 11 Tables

Springer Reference

Editors
Jens Clausen
Institute for Ethics and History of Medicine
University of Tübingen
Tübingen, Germany

Neil Levy
The Florey Institute of Neuroscience and Mental Health
University of Melbourne
Parkville, Australia

ISBN 978-94-007-4706-7 978-94-007-4707-4 (eBook)
ISBN Bundle 978-94-007-4708-1 (print and electronic bundle)
DOI 10.1007/978-94-007-4707-4
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2014946229

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

What Is Neuroethics?

Directly and indirectly, neuroscience touches all of our lives, and its influence can be expected to grow. Mental illness is very common, with a lifetime prevalence of at least 12%, perhaps very much higher (Kessler et al. 2007). The incidence of dementia is growing rapidly, due to our aging population and the fact that more than 30% of all people aged over 80 will suffer from it (Peters 2001). These are problems of the brain/mind, and therefore neuroscience seems to hold out the best hope of understanding them, reducing their incidence, and perhaps even of a cure. Other major health and social problems centrally involve dysfunctions of the brain/mind: think of the enormous problems caused by pathological gambling or drug addiction. Even the worldwide obesity epidemic, which has produced a global population in which there are more obese people than there are undernourished (Macey 2006), might be understood as at least partially a problem that neuroscience can illuminate, insofar as overeating is at least partially to be explained psychologically.

Two major research initiatives in neuroscience funded by the US and the EU, respectively, are expected to increase our knowledge about how the brain works, our understanding of diseases and how to cure or at least treat them and thereby accelerate the influence of neuroscientific findings on our lives. The European Human Brain Project (HBP) addresses seven challenges of neuroscience and ultimately aims at computationally simulating the brain (Markram 2013). The US-funded Brain Activity Map (BAM) project seeks "to fill the gap in our knowledge of brain activity at the circuit level, a scale between single neuron and whole brain function" (Alivisatos et al. 2013).

Since the brain is embedded in and interacts with a body, and humans act in relation to other human beings, the neuro-essentialist notion that the brain defines who we are might seem a bit far-fetched. However, there is no doubt that the brain is the biological substrate of central human characteristics like consciousness and morality. Indeed, all human behavior must be understood as the behavior of creatures with minds; there is no aspect of our lives that neuroscience cannot, in principle, help to illuminate. Neuroscience promises to cure disease and perhaps to help us to achieve our goals at a lower cost, but it also promises to help us to understand the kind of creatures we are. By shedding light on the brain, it illuminates our prized rationality, our creativity, our capacity to produce and appreciate art, even our capacity for awe and transcendence. Some people find the prospect tantalizing and attractive; others are fearful that in explaining we might explain

away. Perhaps we will reveal ourselves to be mere automatons, they worry, rather than beings with a dignity.

Neuroscience also holds out the promise of using its techniques to improve ourselves. The pharmaceuticals developed for the treatment of disease may be used by people who are not ill, to make them "better than well" (Elliot 2003). This, too, is a prospect that excites and appeals to people in equal measures. Some see the promise of an exciting future of broader horizons and a technological utopia, others recall the warnings of *Brave New World* and other dystopias.

To raise these questions is already to be doing neuroethics. Neuroethics is systematic and informed reflection on and interpretation of neuroscience, and related sciences of the mind (psychology in all its many forms, psychiatry, artificial intelligence, and so on), in order to understand its implications for human self-understanding and the perils and prospects of its applications.

Neuroethics has developed as a response to the increasing power and pervasiveness of the sciences of the mind. It has been known for centuries that mental function and dysfunction are closely related to neural function and dysfunction (even Rene Descartes, the 17th century philosopher who is now much derided by neuroscientists for his theory that mind was immaterial, made the connection between brain and mind a central part of his theory). Our knowledge of the nature of the relationship came largely from post-mortem studies of the brains of people known to have particular cognitive problems: Thus, areas responsible for linguistic processing were identified in the 19th century. But it is only recently, with the advent of non-invasive means of studying the living human brain (and especially with the development of functional magnetic resonance imaging, which enables the real-time study of the brain as the person engages in particular tasks), that our knowledge has really begun to expand rapidly. Today the Society for Neuroscience has nearly 42,000 members, all of whom actively working in neuroscience, and holds an annual conference attended by more than 30,000 delegates. There are more than 220 journals dedicated to neuroscience; around 25,000 papers on the brain are published annually. Our knowledge of the brain, and therefore of ourselves, grows rapidly, and with it our powers to intervene in the mind. Neuroethics is at once fascinating and urgent.

Neuroethics is commonly held to have two branches, the *ethics of neuroscience* and the *neuroscience of ethics* (Roskies 2002). Under the first heading, neuroethics is concerned not only with ethical issues in the practice of neuroscience (ethical issues in subject recruitment, in the conduct of neurosurgery, in the reporting of neuroscientific findings in academic journals and the popular press, and so on), but also with ethical issues in the application of neuroscience and the technologies and techniques it helps to develop, inside and outside the clinic. Under this heading, therefore, fall concerns about the use of psychopharmaceuticals, or other techniques (direct current stimulation or implantable electrodes, say) to treat mental illness or to enhance the capacities of those without a diagnosable illness. By the neuroscience of ethics Roskies meant, principally, the ways in which neuroscience might help us to understand morality itself: the principles by which we reason, the relative contribution of emotional and non-emotional processes to moral thought, and

perhaps even the extent to which moral thought sometimes goes wrong. Above, we suggested that neuroethics should not be identified with reflection on neuroscience alone, but be expanded to include reflection on the other sciences of the mind. Correlatively, we suggest that the neuroscience of ethics should also be understood broadly, encompassing not only the ways in which the science of the mind can help us to understand moral reasoning, but also the ways in which it might help us to understand other perennial philosophical issues (the nature of knowledge, the ways in which self-control is exercised and how it may be lost, free will and the mind/brain, and so on). This is, in practice, how neuroethics has been conducted in the past, and it is this broad range of issues that are canvassed in this handbook.

If neuroethics is not to be overwhelmed by the hype that characterizes too much of the popular coverage of neuroscience, it must have a strong and realistic grasp on what is actually possible, on the nature of the brain and its relationship to the mind, and on how best to understand neuroscientific work. The volume therefore begins by canvassing the philosophical foundations of neuroscience, while another section covers the powers and limitations of neuroimaging, our major source of evidence concerning the brain.

Jens Clausen, Neil Levy
Tübingen and Oxford, August 2014

References

Alivisatos, A.P., Chun, M., Church, G.M. et al. 2013. The Brain Activity Map. *Science* 339: 1284–1285.

Elliot C. 2003. *Better than Well*, New York. WW Norton.

Kessler, R.C., Angermeyer, M., Anthony, J. et al. 2007. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* 6: 168–176.

Markram, H. 2013. Seven Challenges for Neuroscience. *Functional Neurology* 28(3): 145–151.

Macye, R. 2006. More fat people in world than there are starving, study finds. *Sydney Morning Herald*, August 15.

Peters, R. 2001. The prevention of dementia. *Journal of Cardiovascular Risk* 8: 253–6.

Roskies, A. 2002. Neuroethics for the New Millenium. *Neuron* 35: 21–23.

About the Editors

Jens Clausen, Institute for Ethics and History of Medicine, University of Tübingen, Tübingen, Germany

Jens Clausen, Dr. rer. nat. (PhD), is professor at the Institute for Ethics and History of Medicine, University of Tübingen, head of the neuroethics group, and managing director of the Clinical Ethics Committee of the University Hospital Tübingen. He also is a member of the Center for Integrative Neuroscience (CIN) and the research ethics commission. Prof. Clausen studied biology and philosophy in Tübingen and was junior researcher and member of the interdisciplinary network project The Status of the Extracorporeal Embryo at the Centre for Ethics and Law in Medicine, Freiburg, Germany, and the Chair for Ethics in Life Sciences, Tübingen. From 2004 to January 2008 he headed the Junior Research Group, Human Nature and Its Relevance in Biomedical Ethics, at the University of Freiburg.

He is guest editor of a double special issues on ethical aspects of neurotechnologies published by Springer in the journal *Neuroethics* (Vol. 6, No. 3, 2013). He has published, edited, or co-edited eight books and is author of more than 50 papers in renowned journals including *Nature*, *European Journal of Neuroscience*, *Current Opinion in Psychiatry, and Neuroethics*.

Neil Levy, The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, Australia

Associate Professor Neil Levy is an Australian Research Council Future Fellow, based at the Florey Institute of Neuroscience and Mental Health, Australia. He is also director of research at the Oxford Centre for Neuroethics. Assoc. Prof. Levy is a philosopher with very wide-ranging interests. He has published more than 100 papers, mainly on applied ethics, free will, philosophy of mind and psychology, as well as continental philosophy, political philosophy, and other topics. He is the author of seven books, include *Neuroethics* (Cambridge University Press, 2007), *Hard Luck* (Oxford University Press [OUP], 2011), and *Consciousness and Moral Responsibility* (OUP, 2014). In 2009, he was awarded the Australia Museum Eureka Award for Research in Ethics.

Section Editors

Françoise Baylis Faculty of Medicine, Novel Tech Ethics, Dalhousie University, Halifax, NS, Canada

Adrian Carter The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia

Jens Clausen Institute for Ethics and History of Medicine, University of Tübingen, Tübingen, Germany

Peggy DesAutels Department of Philosophy, University of Dayton, Dayton, OH, USA

Juan F. Domínguez D. Experimental Neuropsychology Research Unit, School of Psychological Sciences, Monash University, Melbourne, VIC, Australia

Heiner Fangerau Department of History, Philosophy and Ethics of Medicine, University of Ulm, Ulm, Germany

Martha J. Farah Center for Neuroscience & Society, University of Pennsylvania, Philadelphia, PA, USA

Bert Gordijn Institute of Ethics, Dublin City University, Dublin, Ireland

Wayne D. Hall Hall The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia

Sven Ove Hansson Division of Philosophy, Royal Institute of Technology (KTH), Stockholm, Sweden

Hanfried Helmchen Department of Psychiatry & Psychotherapy, Charite´ – University Medicine Berlin, CBF, Berlin, Germany

Anne J. Jacobson University of Houston Center for Neuro-Engineering and Cognitive Science, Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA

Neil Levy Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, VIC, Australia

Reinhard Merkel Faculty of Law, University of Hamburg, Hamburg, Germany

Gualtiero Piccinini Department of Philosophy, University of Missouri – St. Louis, St. Louis, MO, USA

Andrew Pinsent Ian Ramsey Centre for Science and Religion, University of Oxford, Oxford, UK

Eric Racine Neuroethics Research Unit, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

Department of Medicine and Department of Social and Preventive Medicine, Université de Montréal, Montréal, QC, Canada

Departments of Neurology and Neurosurgery, Experimental Medicine & Biomedical Ethics Unit, McGill University, Montréal, QC, Canada

Adina L. Roskies Department of Philosophy, Dartmouth College, Hanover, NH, USA

Stephan Schleim Faculty of Behavioral and Social Sciences, Theory and History of Psychology, Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

Edward H. Spence Centre for Applied Philosophy and Public Ethics (an Australian Research Council Special Research Centre), Charles Sturt University, Canberra, Australia

3TU Centre for Ethics and Technology, University of Twente, Enschede, The Netherlands

Frank W. Stahnisch Department of Community Health Sciences and Department of History, Hotchkiss Brain Institute/Institute for Public Health, The University of Calgary, Calgary, AB, Canada

Jeremy Sugarman Johns Hopkins University, Berman Institute of Bioethics, Baltimore, MD, USA

Matthis Synofzik Department of Neurology, Centre for Neurology and Hertie-Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

Marcos Tatagiba Department of Neurosurgery, Eberhard-Karls University Hospital, Tübingen, Germany

Gerald Walther Division of Peace Studies, University of Bradford, Bradford, UK

Contents

Volume 1: Conceptual Aspects of Neuroethics

Volume 2: Special Issues in Neuroethics

Contributors

Marcus André Acioly Division of Neurosurgery, Fluminense Federal University, Niterói, Rio de Janeiro, Brazil

Division of Neurosurgery, Andaraí Federal Hospital, Rio de Janeiro, Brazil

Department of Neurosurgery, Eberhard–Karls University Hospital, Tübingen, Germany

Kenneth Aizawa Rutgers University - Newark, Newark, NJ, USA

Santiago Amaya Department of Philosophy, Universidad de los Andes, Bogotá, Colombia

Holly Andersen Philosophy, Simon Fraser University, Burnaby, BC, Canada

Daniel Ansari Numerical Cognition Laboratory, Department of Psychology & Brain and Mind Institute, The University of Western Ontario, London, ON, Canada

Mark Balaguer Department of Philosophy, California State University, Los Angeles, CA, USA

Jean-Gaël Barbara Université Pierre et Marie Curie, CNRS UMR 7102, Paris, France

Université Paris Diderot, CNRS UMR 7219, Paris, France

Imre Bárd London School of Economics and Political Science, London, UK

Simon Baron-Cohen Autism Research Centre, Psychiatry Department, Cambridge University, Cambridge, UK

Robert Bauer Translational and Functional & Restorative Neurosurgery, Department of Neurosurgery, University Hospital Tübingen, Tübingen, Germany

International Centre for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany

Françoise Baylis Faculty of Medicine, Novel Tech Ethics, Dalhousie University, Halifax, NS, Canada

Tim Bayne Philosophy, School of Social Sciences, University of Manchester, Manchester, UK

Paul Biegler Centre for Human Bioethics, School of Philosophical, Historical and International Studies Monash University Faculty of Arts, Clayton, VIC, Australia

Niels Birbaumer Institute of Medical Psychology and Behavioral Neurobiology, Tübingen, Germany

IRCCS, Ospedale San Camillo, Istituto di Ricovero e Cura a Carattere Scientifico, Lido di Venezia, Italy

Robyn Bluhm Department of Philosophy and Religious Studies, Old Dominion University, Norfolk, VA, USA

Hilary Bok Johns Hopkins University, Baltimore, MD, USA

Christoph Bublitz Faculty of Law, University of Hamburg, Hamburg, Germany

Alena Buyx Centre for Advanced Studies in Bioethics, University Hospital Münster and University of Münster, Emmy Noether Research Group Bioethics and Political Philosophy, Münster, Germany

School of Public Policy, University College London, London, UK

Danilo Bzdok Institut für Neurowissenschaften und Medizin (INM-1), Jülich, Germany

Benjamin Capps Centre for Biomedical Ethics, National University of Singapore Yong Loo Lin School of Medicine, Singapore

Adrian Carter The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia

Jennifer Chandler Faculty of Law, Common Law, University of Ottawa, Ottawa, ON, Canada

Paul P. Christopher Department of Psychiatry & Human Behavior, Alpert Medical School, Brown University, Providence, RI, USA

Jens Clausen Institute for Ethics and History of Medicine, University of Tübingen, Tübingen, Germany

Malcolm Dando Division of Peace Studies, School of Social and International Studies, University of Bradford, Bradford, UK

Athena Demertzi Coma Science Group, Cyclotron Research Center & Neurology Department, University of Liège, Liège, Belgium

Peggy DesAutels Department of Philosophy, University of Dayton, Dayton, OH, USA

Juan F. Domínguez D. Experimental Neuropsychology Research Unit, School of Psychological Sciences, Monash University, Melbourne, VIC, Australia

John M. Doris Philosophy-Neuroscience-Psychology Program and Philosophy Department, Washington University, St. Louis, MO, USA

Thomas Douglas Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Oxford, UK

Martin Dresler Max Planck Institute of Psychiatry, Munich, Germany

Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands

Laura B. Dunn Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA

Simon B. Eickhoff Institut für Neurowissenschaften und Medizin (INM-1), Jülich, Germany

Heiner Fangerau Department of History, Philosophy and Ethics of Medicine, University of Ulm, Ulm, Germany

Martha J. Farah Center for Neuroscience & Society, University of Pennsylvania, Philadelphia, PA, USA

G. C. Feigl Department of Neurosurgery, University Hospital Tübingen, Tübingen, Germany

Fiona Fidler Australian Centre of Excellence for Risk Analysis (ACERA), Environmental Science, School of Botany, University of Melbourne, Carlton, VIC, Australia

Cordelia Fine Melbourne School of Psychological Sciences & Melbourne Business School & Centre for Ethical Leadership, University of Melbourne, Carlton, VIC, Australia

Joseph J. Fins Division of Medical Ethics, New York Presbyterian–Weill Cornell Medical Center, Weill Medical College of Cornell University, New York, NY, USA

Consortium for the Advanced Study of Brain Injury (CASBI), Weill Cornell Medical College & Rockefeller University, New York, NY, USA

Farah Focquaert Bioethics Institute Ghent, Department of Philosophy and Moral Sciences, Ghent University, Ghent, Belgium

Paul Foley Unit for History and Philosophy of Science, University of Sydney, Sydney, NSW, Australia

Neuroscience Research Australia, Randwick, Sydney, NSW, Australia

Paul J. Ford Department of Bioethics, NeuroEthics Program, Cleveland Clinic, Cleveland, OH, USA

Cynthia Forlini Institut de recherches cliniques de Montréal (IRCM), Neuroethics Research Unit, Montréal, QC, Canada

UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia

Thorsten Galert German Reference Centre for Ethics in the Life Sciences, Bonn, Germany

Delia Gavrus Department of History, University of Winnipeg, Winnipeg, MB, Canada

Alireza Gharabaghi Translational and Functional & Restorative Neurosurgery, Department of Neurosurgery, University Hospital Tübingen, Tübingen, Germany

International Centre for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany

Jay N. Giedd National Institutes of Health, National Institute of Mental Health, Bethesda, MD, USA

Carl Gillett Northern Illinois University, DeKalb, IL, USA

James Giordano Neuroethics Studies Program, Pellegrino Center for Clinical Bioethics, Division of Integrative Physiology; Department of Biochemistry and Integrative Program in Neurosciences, Georgetown University Medical Center, Washington, DC, USA

Human Science Center, Ludwig Maximilians Universität, Munich, Germany

Walter Glannon Department of Philosophy, University of Calgary, Calgary, AB, Canada

Andrea L. Glenn Center for Prevention of Youth Behavior Problems, Department of Psychology, The University of Alabama, Tuscaloosa, AL, USA

Imogen Goold Faculty of Law, University of Oxford, St Anne's College, Oxford, UK

Bert Gordijn Institute of Ethics, Dublin City University, Dublin, Ireland

Adam Green Department of Philosophy, Azusa Pacific University, Azusa, CA, USA

Dominik Groß Medical School, RWTH Aachen University, Aachen, Germany

Wayne D. Hall The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia

Queensland Brain Institute, The University of Queensland, St Lucia, QLD, Australia

Sven Ove Hansson Division of Philosophy, Royal Institute of Technology (KTH), Stockholm, Sweden

Chelsea Helion Department of Psychology, Cornell University, Ithaca, NY, USA

Hanfried Helmchen Department of Psychiatry & Psychotherapy, Charité – University Medicine Berlin, CBF, Berlin, Germany

Spencer Phillips Hey Biomedical Ethics Unit, McGill University, Montreal, QC, Canada

Sam Horng Department of Neurology, Mount Sinai Medical Center, New York, NY, USA

Merv Hyde Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Sippy Downs, QLD, Australia

Anne J. Jacobson University of Houston Center for Neuro-Engineering and Cognitive Science, Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA

Karim Jebari Department of Philosophy, Royal Institute of Technology (KTH), Stockholm, Sweden

Catherine Jefferson Department of Social Science, Health & Medicine, King's College London, London, UK

Sara B. Johnson Johns Hopkins University School of Medicine, Baltimore, MD, USA

Josephine Johnston The Hastings Center, Garrison, NY, USA

Fabrice Jotterand Institute for Biomedical Ethics, University of Basel, Basel, Switzerland

David Michael Kaplan Department of Cognitive Science, Macquarie University, Sydney, NSW, Australia

Antti Kauppinen Trinity College, Dublin, Ireland

Michael L. Kelly Department of Neurosurgery, Cleveland Clinic, Cleveland, OH, USA

Jeanette Kennett Department of Philosophy, Macquarie University, Sydney, NSW, Australia

Jonathan Kimmelman Studies for Translation, Ethics and Medicine (STREAM), Biomedical Ethics/Social Studies of Medicine/Department of Human Genetics, McGill University, Montreal, QC, Canada

Linda Komesaroff Deakin University, Waurn Ponds, VIC, Australia

Paul A. Komesaroff Monash Centre for Ethics in Medicine and Society, Monash University, Caulfield East, VIC, Australia

Norbert Konrad Institut für Forensische Psychiatrie, Charité – Universitätsmedizin, Berlin, Germany

Timothy Mark Krahn Novel Tech Ethics, Dalhousie University, Halifax, NS, Canada

William Langley University of Houston Center for Neuro-Engineering and Cognitive Science, Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA

Charles D. Laughlin Department of Sociology and Anthropology, Carleton University, Ottawa, ON, Canada

Steven Laureys Coma Science Group, Cyclotron Research Center & Neurology Department, University of Liège, Liège, Belgium

Bartlomiej A. Lenart Department of Philosophy, University of Alberta, Edmonton, AB, Canada

Christian Lenk Ulm University, Ulm, Germany

Neil Levy Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, VIC, Australia

Anne Lingford-Hughes Centre for Neuropsychopharmacology, Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK

Jayne Lucke University of Queensland Centre for Clinical Research, The University of Queensland, Brisbane, QLD, Australia

R. Luerding Department of Neurology, University of Regensburg Medical Center, Regensburg, Germany

Catriona Mackenzie Department of Philosophy, Macquarie University, Sydney, NSW, Australia

Corey J. Maley Department of Philosophy, Princeton University, Princeton, NJ, USA

Hannah Maslen Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK

Paul H. Mason Woolcock Institute of Medical Research, Glebe, NSW, Australia

Debra J. H. Mathews Johns Hopkins Berman Institute of Bioethics, Baltimore, MD, USA

Steve Matthews Plunkett Centre for Ethics (St Vincent's and Mater Health Sydney), Department of Philosophy, Australian Catholic University, Sydney, NSW, Australia

Donatella Mattia Clinical Neurophysiology, Neuroelectrical Imaging and BCI Laboratory, Fondazione Santa Lucia IRCCS, Rome, Italy

Thomas R. McCormick Department Bioethics and Humanities, School of Medicine, University of Washington, Seattle, WA, USA

Iain McGilchrist The Bethlem Royal and Maudsley Hospital, London, UK

Reinhard Merkel Faculty of Law, University of Hamburg, Hamburg, Germany

M. Milian Department of Neurosurgery, University Hospital Tübingen, Tübingen, Germany

Franklin G. Miller Department of Bioethics, Clinical Center, National Institutes of Health, Bethesda, MD, USA

Masamichi Minehata Division of Peace Studies, University of Bradford, Bradford, UK

Sabine Müller Department for Psychiatry and Psychotherapy CCM, Charité – Universitätsmedizin, Berlin, Germany

Liam Nestor Centre for Neuropsychopharmacology, Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK

Tatyana Novossiolova Bradford Disarmament Research Centre, Division of Peace Studies, University of Bradford, Bradford, UK

Justin Oakley Centre for Human Bioethics, School of Philosophical, Historical and International Studies Monash University Faculty of Arts, Clayton, VIC, Australia

Fiachra O'Brolcháin Institute of Ethics, Dublin City University, Dublin, Ireland

Elisabeth Pacherie Institut Jean Nicod – UMR 8129, ENS, EHESS, CNRS, Paris, France

Anna Pacholczyk Centre for Social Ethics and Policy, School of Law, University of Manchester, Manchester, UK

Institute for Science Ethics and Innovation, Faculty of Life Sciences, University of Manchester, Manchester, UK

Erik Parens The Hastings Center, Garrison, NY, USA

Brad Partridge University of Queensland Centre for Clinical Research, The University of Queensland, Brisbane, QLD, Australia

Norbert W. Paul History, Philosophy, and Ethics of Medicine, Johannes Gutenberg University Medical Center, Mainz, Germany

Gualtiero Piccinini Department of Philosophy, University of Missouri – St. Louis, St. Louis, MO, USA

Andrew Pinsent Ian Ramsey Centre for Science and Religion, University of Oxford, Oxford, UK

David A. Pizarro Department of Psychology, Cornell University, Ithaca, NY, USA

Eric Racine Neuroethics Research Unit, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

Department of Medicine and Department of Social and Preventive Medicine, Université de Montréal, Montréal, QC, Canada

Departments of Neurology and Neurosurgery, Experimental Medicine & Biomedical Ethics Unit, McGill University, Montréal, QC, Canada

Adrian Raine Departments of Criminology, Psychiatry, and Psychology, Jerry Lee Center of Criminology, University of Pennsylvania, Philadelphia, PA, USA

Stephen Reyna Max Planck Institute of Social Anthropology, Halle, Germany

Regina A. Rini University of Oxford, Oxford, UK

Sabine Rona Department of Neurosurgery, University Hospital, Eberhard Karls University, Tübingen, Germany

Diana Rose Health Service and Population Research Department, King's College London Institute of Psychiatry, London, UK

Adina L. Roskies Department of Philosophy, Dartmouth College, Hanover, NH, USA

Chiara Saviane Interdisciplinary Laboratory for Advanced Studies, Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy

Marya Schechtman University of Illinois at Chicago, Chicago, IL, USA

Maartje Schermer Department Medical Ethics and Philosophy of Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

Stephan Schleim Faculty of Behavioral and Social Sciences, Theory and History of Psychology, Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

Neurophilosophy, Munich Center for Neurosciences, Ludwig–Maximilians–University Munich, Munich, Germany

Elizabeth Shaw School of Law, University of Aberdeen, Aberdeen, UK

Ilina Singh Department of Social Science, Health & Medicine, King's College London, London, UK

Anke Snoek Philosophy Department, Macquarie University, Sydney, NSW, Australia

Surjo R. Soekadar Department of Psychiatry and Psychotherapy, Applied Neurotechnology Lab/Institute of Medical Psychology and Behavioral Neurobiology, University Hospital Tübingen, Tübingen, Germany

Tamler Sommers University of Houston, Houston, TX, USA

Edward H. Spence Centre for Applied Philosophy and Public Ethics (an Australian Research Council Special Research Centre), Charles Sturt University, Canberra, Australia

3TU Centre for Ethics and Technology, University of Twente, Enschede, The Netherlands

Frank W. Stahnisch Department of Community Health Sciences and Department of History, Hotchkiss Brain Institute/Institute for Public Health, The University of Calgary, Calgary, AB, Canada

Jeremy Sugarman Johns Hopkins University, Berman Institute of Bioethics, Baltimore, MD, USA

Jacqueline Sullivan Department of Philosophy and Rotman Institute of Philosophy, University of Western Ontario, London, ON, Canada

Matthis Synofzik Department of Neurology, Centre for Neurology and Hertie-Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

George Szmukler King's College London, Institute of Psychiatry, London, UK

Guglielmo Tamburrini DIETI – Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Università di Napoli Federico II, Naples, Italy

Marcos Tatagiba Department of Neurosurgery, Eberhard-Karls University Hospital, Tübingen, Germany

Michael Trimble Institute of Neurology, London, UK

Odile Nogueira Ugarte Division of Neurosurgery, Andaraí Federal Hospital, Rio de Janeiro, Brazil

Postgraduation Program in Neurology, Federal University of the State of Rio de Janeiro, Rio de Janeiro, Brazil

Yvette van der Eijk Centre for Biomedical Ethics, National University of Singapore Yong Loo Lin School of Medicine, Singapore

Patrick Vargas Department of Advertising, University of Illinois at Urbana-Champaign, IL, USA

Fernando Vidal ICREA (Catalan Institution for Research and Advanced Studies), CEHIC (Center for the History of Science, Autonoma University of Barcelona), Barcelona, Spain

Unitat d'Història de la Medicina, Facultat de Medicina, M6/130, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain

Nicole A. Vincent Philosophy Department, Georgia State University, Atlanta, GA, USA

Philosophy Section, Technische Universiteit Delft, Delft, The Netherlands

Aku Visala Department of Anthropology, University of Notre Dame, Notre Dame, IN, USA

Faculty of Theology, University of Helsinki, Helsinki, Finland

Mary Walker Department of Philosophy, Macquarie University, Sydney, NSW, Australia

Gerald Walther Division of Peace Studies, University of Bradford, Bradford, UK

Robert A. Wilson Department of Philosophy, University of Alberta, Edmonton, AB, Canada

Murat Yücel Monash Clinical and Imaging Neuroscience, School of Psychology and Psychiatry, Monash University, Melbourne, VIC, Australia

Section I

Foundational Issues in Cognitive Neuroscience

Foundational Issues in Cognitive Neuroscience: Introduction

1

Gualtiero Piccinini

Contents

Abstract

Cognitive neuroscience raises several foundational issues. A first issue is how to account for our feeling that we are in control of our actions. A second issue is how to account for the relation between the mind and the nervous system. A third issue is how cognitive neuroscience supports its conclusions. A fourth issue is how cognitive neuroscience explains phenomena. A fifth issue is how to account for the notions of neural representation and neural computation.

Is the Mind Causally Efficacious in the Physical World?

To many people, the mind appears to be very different in nature from the physical stuff of which nervous systems are made. The mind seems to be made of *mental* stuff, or *experience*, stuff that strikes many people as . . . fundamentally different from physical stuff. Perhaps the mind is a nonphysical substance, or at least a collection of nonphysical properties that attaches to the nervous system. But if

G. Piccinini
Department of Philosophy, University of Missouri – St. Louis, St. Louis, MO, USA
e-mail: piccininig@umsl.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_3,

the mind is fundamentally different from physical substances and properties, it becomes unclear how mental states could possibly be caused by physical stimuli (as in perception) or cause physical responses (as in behavior). This is the problem of mental causation, which is discussed in ▶ Chap. 5, “Mental Causation” by Holly Anderson.

One view is that mental stuff, no matter how fundamentally different it is from physical stuff, can causally interact with physical stuff. This is called *interactionist dualism*. Interactionist dualism has the advantage of vindicating our feeling that our mind interacts with the physical world, but it has the remarkable disadvantage of making it mysterious how this occurs. No one has ever explained how something nonphysical could possibly interact with something physical. In addition, interactionist dualism implies that cognitive neuroscience cannot explain cognition and behavior—the true explanation lies in the nonphysical (mental) stuff, which presumably is not accessible to standard neuroscientific methods.

Another view is that mental stuff is causally inert after all—or at least it does not causally interact with the *physical* stuff of which our nervous systems are made, or if it is *caused by* physical events, it *causes no* physical events of its own. Maybe mental states interact with each other, but they never interact with anything physical. This view—called epiphenomenalism—frees cognitive neuroscience to explain our behavior in terms of neural mechanisms and processes, and some cognitive neuroscientists have endorsed it. Unfortunately, epiphenomenalism also implies that our mind has nothing to do with our behavior, which many people find hard to swallow.

The last possibility is that the mind is *physical*—presumably, the mind is some aspect of the nervous system and its activity. Those who think the mind is fundamentally different from physical stuff are just wrong. This view is called *physicalism*. Like interactionist dualism, physicalism vindicates our feeling that our mind interacts with the physical world. Like epiphenomenalism, physicalism implies that cognitive neuroscience can explain cognition and behavior. But physicalism raises new questions as well.

How Does the (Physical) Mind Relate to the Nervous System?

Even if the mind is physical in a broad sense (physicalism), there remains the question of how, more precisely, it is metaphysically related to the nervous system. This issue is discussed in the entry by Aizawa and Gillett.

One possibility is that mental properties are just physical properties of the nervous system. This view is called *reductionism*. All that cognitive neuroscience has to do in order to explain cognition and behavior is to find the physical properties of the nervous system that are identical to the mental properties. Reductionism has an appealing simplicity, but—at least in its strongest forms—it has a couple of disadvantages as well. First, (strong) reductionism does not sit well with the prevalent explanatory strategy within cognitive neuroscience, which is mechanistic (see below); this is because fitting a phenomenon within a multilevel mechanistic explanation involves both showing how the phenomenon is produced by a series of

organized components, which sounds reductionist, *and* showing how the phenomenon contributes to a higher-level mechanism, which sounds antireductionist. Second, (strong) reductionism does not sit easily with the plausible view that cognition may be physically realizable by systems that possess no nervous system—for instance, silicon-based robots.

Another possibility is that mental properties are *realized* by physical properties of the nervous system without being identical to them. According to this view, cognitive neuroscience may still explain cognition and behavior by finding the physical properties of the nervous system that realize mental properties, but this leaves open the possibility that other physical systems—say, silicon-based robots—also realize the same mental properties in some physically different way. This view also seems to fit well the mechanistic explanatory style that prevails in neuroscience.

A third possibility is that mental properties are higher-level physical properties that emerge in nervous systems in addition to their lower-level physical properties—emergent properties are novel and irreducible to lower-level properties. This is called (strong) emergentism. Emergentism vindicates the common feeling that there is something special about our mind, but it introduces a mystery about how the emergent properties emerge. In addition, the emergent properties appear ontologically redundant: once the lower-level properties have done their causal work, there does not seem to be any causal work left over for the emergent properties to do. If this causal exclusion argument is correct, emergentism collapses back into epiphenomenalism—the view that denies the mind any influence on the physical world.

How Does Cognitive Neuroscience Support Its Conclusions?

Cognitive neuroscientists intervene on the nervous system, collect data, and draw conclusions. What this means and the epistemic risks involved are discussed in the entry by Jacqueline Sullivan.

A cognitive neuroscience experiment requires subjects to engage in a cognitive task. After that, behavior and neural activity are recorded using a variety of techniques that include both neuroimaging methods and neurophysiological recording. Computational models may also be used to understand how subjects may be able to process solve the given task. A prediction is made—for instance, about whether a certain area of the brain is involved in a cognitive task. In the best-case scenario, the data collected will adjudicate between these three competing hypotheses and point to the one that is best supported by the data.

The data from an experiment will serve their function only to the extent that they are reliable and valid. Data are *reliable* just in case they are unlikely to support false conclusions. Data are *valid* just in case they support the conclusions that they were intended to support. The intended conclusion may be about human cognition in the wild, whereas the data may be collected from rats or fruit flies operating in a constrained laboratory environment. The extrapolation from the latter to the former carries inductive risk.

Imaging techniques face a specific set of epistemic challenges, which must be handled appropriately. First, the experimental paradigms used in conjunction with imaging technology may or may not be sufficient to individuate the cognitive function under investigation. Second, typical neuroimaging techniques do not measure neural activity directly, but some correlate of it, which raises the question of what the exact correlation is between the two. Third, neuroimaging data must be processed before they are usable to draw any conclusion, and this data processing may introduce mistakes and biases that affect the reliability and validity of the conclusions. In summary, experiments in cognitive neuroscience are fraught with epistemic risks, which we should consider when evaluating the conclusions that are drawn from them.

How Does Cognitive Neuroscience Explain?

There is an old idea about scientific explanation: it consists of deriving a phenomenon from the laws of nature together with initial conditions. The entry by Kaplan discusses why this old idea is poorly suited for explanation in cognitive neuroscience and proposes a replacement. For starters, cognitive neuroscience rarely discovers or invokes anything resembling laws of nature. A better account is that cognitive neuroscience explains by providing multilevel mechanisms.

A mechanism is an organized collection of components, each of which plays a role in producing a phenomenon. A mechanistic explanation explains a phenomenon by showing how the roles played by each component, when the components are organized in the appropriate way, produce the phenomenon. Each role of each component of a mechanism is a phenomenon of its own, which may be explained mechanistically by going down one level and looking at the component's sub-components, the roles they play, and the way they are organized. By the same token, a phenomenon produced by a mechanism may play a role in a larger mechanism, whose behavior may be explained mechanistically by going up one level and looking at what a mechanism contributes to the larger system that contains it as a component.

Cognitive neuroscience explains cognition and behavior in terms of multilevel neural mechanisms. If we start from cognition and behavior, those are phenomena to be explained in terms of neural systems (cortical areas, cerebellum, brainstem, etc.), their roles, and their organization. The roles played by neural systems, in turn, are phenomena to be explained in terms of their components (cortical columns, nuclei, etc.), their roles, and their organization. Going down to even lower levels, we find neural networks, neurons, and their components. If and when we understand cognition and behavior at all these levels, we will have a multilevel mechanistic explanation of cognition and behavior.

The above picture assimilates explanation of cognition and behavior to explanation in other mechanistic sciences, such as other areas of physiology and engineering. Is there anything that distinguishes neurocognitive explanation from other mechanistic explanations?

How Does the Nervous System Think?

An important part of the explanation for how nervous systems manage to think is that they collect information from the organism and the environment, use that information to construct representations, and perform computations on such representations. Nervous systems have the distinctive function of performing computations over representations in order to control the organism—that sets them apart from most other mechanisms. The entry by Maley and Piccinini elaborates on this point.

To begin with, nervous systems possess receptors that respond to a wide variety of physical stimuli: light, sounds, odors, pressure, and more. These receptors transduce these physical stimuli into spike trains—sequences of neuronal signals that encode information about the physical stimuli. Neural signals are transmitted to the central nervous system, where they are processed. Such processing—called neural computations—extracts information that is implicit in the signals and combines such information with internal information about the organism's needs, beliefs, and desires—all of which are also encoded as states of the central nervous system. The outputs of neural computations are updated internal states as well as motor outputs—the behaviors of the organism.

This is how cognitive neuroscience explains cognition and behavior—as the outcome of computations over representations performed by multilevel neural mechanisms. Or at least, this is how current cognitive neuroscience explains some aspects of cognition and behavior.

Conclusion and Future Directions

Cognitive neuroscience is an exciting field in the middle of a growth spurt. It collects evidence at multiple levels of mechanistic organization by performing sophisticated experiments using innovative techniques. It integrates such evidence from multiple levels into multilevel mechanistic explanations. It explains cognitive phenomena mechanistically in terms of neural computations operating on neural representations.

There remains room for debate and further work on how to understand cognitive neuroscience, the explanations it gives, their relation to psychological explanations, and the role of the mind in physical world.

Cross-References

► Experimentation in Cognitive Neuroscience and Cognitive Neurobiology
► Explanation and Levels in Cognitive Neuroscience
► Mental Causation
► Neural Representation and Computation
► Realization, Reduction, and Emergence: How Things Like Minds Relate to Things Like Brains

Explanation and Levels in Cognitive Neuroscience

2

David Michael Kaplan

Contents

Abstract

Talk of levels is widespread in the life sciences including neuroscience. This chapter explores some of the most common characterizations of levels found in neuroscience, with a specific focus on how levels of organization are invoked in explanations.

D.M. Kaplan
Department of Cognitive Science, Macquarie University, Sydney, NSW, Australia
e-mail: david.kaplan@mq.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_4,

Introduction

Talk of levels is prevalent in biology and neuroscience. Investigators in the life sciences routinely deploy various notions of levels including levels of organization, levels of complexity, levels of mechanisms, levels of processing, levels of description, and levels of abstraction. Some of these notions are terminological variants of one another. Others play importantly distinct roles. Some are ontological, describing aspects of the phenomena or target systems under investigation. Others are epistemic, referring to features of the theories, generalizations, and models that scientists build in order to describe, predict, and explain those phenomena. This chapter provides an overview of some of the most common characterizations of levels found in neuroscience, with a specific focus on how levels of organization are invoked in relation to explanation. After characterizing the dominant philosophical approaches to thinking about levels and explanation, the relative advantages of embracing the view of levels embodied by the mechanistic perspective are highlighted. Next, several pressing challenges facing the mechanistic approach are surveyed and how they might be met is briefly explored. Finally, the chapter discusses how the mechanistic approach affords a fruitful viewpoint on another, seemingly unrelated, notion of level frequently invoked in neuroscience – that of Marrian levels of analysis.

By way of introduction, consider one prominent notion of level in the biological sciences – *level of biological organization*. Many biologists have emphasized the importance of hierarchical levels of organization in living systems (e.g., Lobo 2008; MacMahon et al. 1978; Novikoff 1945; Woodger 1929). The idea of hierarchical levels of organization is now established in biology and forms a cornerstone of modern biological theory. General biology textbooks (e.g., Sadava et al. 2009) invariably begin by introducing the biological world as consisting of stratified levels arranged hierarchically such that a particular level of biological organization serves as a sublevel for the next higher level: atoms assemble to form molecules, which in turn assemble into macromolecules, which assemble into cells, which assemble into tissues, which assemble into organs, which assemble into systems, which assemble into organisms, which assemble into populations, which assemble into communities, and so on. The idea of distinct levels of organization is canonical in biology.

Many cognitive scientists and neuroscientists similarly acknowledge the relevance of levels and hierarchical organization for understanding complex systems such as computers, networks, brains, and human organizations (e.g., Simon 1996; Churchland and Sejnowski 1992; Findlay and Thagard 2012; Holland 2000; Sporns 2010). For example, Herbert Simon (1996), a pioneer in cognitive science and artificial intelligence, influentially argues that the concept of hierarchy is the key concept in complex systems research. The concepts of hierarchy and levels are tightly linked: levels are positions in hierarchies and hierarchies are ordering schemes in which elements are arranged at different levels according to some principle or criterion. Churchland and Sejnowski (1992) likewise maintain that the notion of levels plays a fundamental role in neuroscience, serving to structure both how research problems are conceived and how neural systems are investigated.

Numerous investigators have also highlighted how comprehending complex systems, especially complex information-processing systems like the brain, demand an approach that is either carried out at the appropriate level of description for the system under investigation (Anderson 1972; Carandini 2012) or successfully integrates multiple levels of analysis or system description (Bechtel 1994; Bermudéz 2005, 2010; Dawson 1998; Marr 1982; Sporns 2010). Strikingly, even one of the stated goals of neuroscience research articulated in the *Society for Neuroscience's* mission statement is to "advance the understanding of the brain and the nervous system. by facilitating the integration of research directed at all levels of biological organization" (http://www.sfn.org/about/mission-and-strategic-plan). Clearly, many brain researchers and cognitive scientists appreciate that the notion of levels comprises a critically important part of their conceptual toolkit.

Finally, philosophers of science have argued that a dominant form of explanation in biology and neuroscience is mechanistic explanation, which inherently exhibits a *multilevel* structure involving entities and activities occupying multiple levels of organization (e.g., Bechtel 2008; Bechtel and Richardson 1993/2010; Craver 2007; Winther 2011).

Despite the ubiquity and operational importance of various concepts of levels in biology and neuroscience, it is not always perfectly clear how these different concepts should be understood. Interestingly, confusion surrounding notions of levels is not a recent phenomenon – it has a long history (e.g., Bunge 1977). Despite considerable philosophical discussion of the topic, open questions remain. The aims of this chapter are to survey several prominent notions of levels and explanation and highlight the advantages of embracing a mechanistic perspective concerning levels in neuroscience.

Basic Distinctions and Defining Questions

Before proceeding, it is important to flag a basic distinction that will be relied upon in what follows. As mentioned above, the term "level" is ambiguous in science – in one context "level" might refer to the relationship between temporal stages in a causal or information-processing sequence, while in another context the term might refer to the relationship between scientific theories. Nevertheless, a basic distinction between *ontological levels* and *epistemic levels* can begin to impose some order on this diversity. In his useful taxonomy of levels, Craver draws an equivalent distinction between *levels of science* and *levels of nature* (Craver 2007, p. 171). Levels of science subsume relationships among the basic *products* or output of scientific research (e.g., models and theories) and the basic *units* or institutional divisions of science (e.g., research fields and disciplines). For example, the proposal that string theory is at a lower level than the theory of genetic drift posits something about the relationship among levels of science. These are *epistemic levels*. By contrast, levels of nature refer to relationships among *phenomena* (e.g., properties, states, events, and objects) in the natural world that are targets of

scientific investigation. For example, the claim that quarks occupy a lower level than DNA asserts something about the relationship among levels of nature. Claims about levels of nature carry ontological commitments. These are *ontological levels*.

One further resource to be recruited in this chapter is the following set of defining questions about levels also outlined by Craver (2007, pp. 163–195):

The relata question: What is the nature of the relata in the proposed relationship between levels?

The ranking question: What makes a given element occupy a different (higher or lower) level than some other element?

The parity question: What makes two (or more) elements occupy the same level?

These questions identify plausible theoretical desiderata that any adequate philosophical account of levels must satisfy. Consequently, they will be used as benchmarks for assessing the various approaches to levels considered in what follows.

Explanation and Levels: The Traditional Approach

The basic logical empiricist account of explanation – the *covering law model* of scientific explanation (Hempel and Oppenheim 1948; Hempel 1965) – and its associated account of theory reduction (Nagel 1961) continue to exert a profound influence on contemporary philosophical discussions of these topics (e.g., Woodward 2003, 2009; Brigandt and Love 2012). Despite well-known limitations, one lasting legacy of these once-dominant accounts is the view of levels they have inspired.

The Covering Law Model of Explanation

The underlying idea of the covering law model is that to explain something is to show how to derive it in a logical argument. Accordingly, a scientific explanation should be readily expressible as a logical argument in which the *explanandum-phenomenon* (that which is being explained) appears as the conclusion of the argument and the *explanans* (that which does the explaining) appears as the premises. The premise set comprising the explanans typically includes statements describing the relevant empirical conditions under which the phenomenon obtains, known as *initial conditions*. One additional requirement of the model is that at least one of the premises must describe a *general law* required for the derivation of the explanandum. The explanans is supposed to provide good evidential grounds for expecting the occurrence of the explanandum-phenomenon. The account can be depicted schematically as follows (Hempel 1965, p. 336):

$$\frac{\begin{array}{c} C_1, C_2, \ldots, C_k \\ L_1, L_2, \ldots, L_r \end{array}}{\therefore E}$$

where $C_1, C_2, \ldots, C_k$ are premises describing initial conditions and $L_1, L_2, \ldots, L_r$ are premises expressing the relevant general laws. In its original formulation (Hempel and Oppenheim 1948), the explanandum-phenomenon E is characterized as a *deductive* consequence of the premises. This specific version of the account has come to be known as the *deductive–nomological* or *D–N model*. Hempel later extended the account to apply to explanations involving the subsumption of events under statistical rather than deterministic laws. According to the *inductive–statistical* or *I–S model* (Hempel 1965), good statistical explanations are ones in which the explanandum-phenomenon is rendered expectable with a high degree of probability based on the premises. The covering law model is typically treated as the general account, subsuming both D–N and I–S style explanations as species (for additional discussion, see Woodward 2009).

In its most general formulation, the covering law account is intended to apply both to the explanation of particular events that are restricted in space and time such as the Space Shuttle Challenger crash or the extinction of the dinosaurs, as well as to the explanation of general regularities or laws such as the explanation of Kepler's laws of planetary motion in terms of more basic laws of Newtonian mechanics (in combination with a specification of initial conditions concerning the distribution of planetary bodies in the solar system) or the explanation of Mendel's laws of inheritance in terms of more fundamental generalizations from molecular genetics. Derivations of one or more sets of laws (comprising a theory) from another set of laws (comprising another theory) are known as *intertheoretic reductions*. Given that the covering law account is designed to apply to the explanatory relationships holding among laws or theories, it should not be too difficult to see how a corresponding account of theory reduction emerges.

The Nagelian Model of Theory Reduction

According to the traditional logical empiricist perspective, explanation and theory reduction are closely linked (For additional discussion of reduction and related topics, see ▶ Chap. 4, "Realization, Reduction, and Emergence: How Things Like Minds Relate to Things Like Brains"). Just as the D–N model ties successful explanation to the presence of a deductive or derivability relationship between explanans and explanandum statements, so too successful reduction is tied to the presence of a deductive relationship between two theories. Intertheoretic reduction is thus naturally treated as a special case of deductive–nomological explanation.

Most discussions of theory reduction in philosophy of science have been heavily influenced by Nagel's model of theory reduction (Nagel 1961, pp. 366–397). One of Nagel's central aims was to elucidate the logical or explanatory relationships between the scientific disciplines and especially the relationships between the theories of those disciplines. Levels talk is quite naturally recruited to characterize his objective. Nagel wanted to provide an account of the conditions that must obtain in order for a theory (or law) from a higher-level science such as psychology to be reduced to a theory (or law) from a lower-level science such as neuroscience or biophysics. The basic answer Nagel defends is that a lower-level theory serves to reduce (and also *explains*) a higher-level theory when it can be used to logically derive the lower-level theory. More formally, in order for a suitably axiomatized theory T_1 (the *reducing theory*) to reduce, and thereby explain, a suitably axiomatized theory T_2 (the *reduced theory*), T_2 must be logically entailed by or derivable from T_1:

$$T_1 \vdash T_2$$

What this means is that all the laws and all of the observational consequences of the reduced theory T_2 can be derived from information contained in the reducing theory T_1. Nagel (1961) argues that the relationship between classical thermodynamics and statistical mechanics cleanly fit this account of reduction. He maintains that the laws of thermodynamics (e.g., the Boyle–Charles law relating temperature, pressure, and volume in a gas) can be derived from – and so are successfully reduced to and explained by – more fundamental laws of statistical mechanics (e.g., laws characterizing the aggregate behavior of the constituent molecules). Since the terminology invoked by the reduced and reducing theories often differs in appreciable ways, the so-called bridge principles are required to establish identities between the terms of the two theories. For example, a bridge principle might link terms from thermodynamics such as "heat" with those of statistical mechanics such as "mean molecular energy." Finally, because the reduced theory will typically only apply over a restricted part of the domain of the reducing theory (Nagel 1961) or at certain limits (Batterman 2002), *boundary conditions* that set the appropriate range for the reduction are often required. For example, the regularities captured by the higher-level gas law of thermodynamics only obtain for a restricted range of systems under a restricted range of conditions (e.g., gases operating within a certain specified range of pressures and temperatures). Without specifying boundary conditions, most derivations of higher-level laws (in the reduced theory) from the lower-level laws (the reducing theory) would be unsuccessful.

Because statistical mechanics provides a lower-level (i.e., molecular) explanation of macroscopic thermodynamic quantities such as work, heat, and entropy, and the regularities these quantities enter into, it is generally understood to provide a clear example of the reduction of a higher-level theory to a lower-level one (for recent discussion, see Dizadji-Bahmani et al. 2010; Callendar 1999; Sklar 1999). Whatever its successes might be, however, Nagel's model of reduction provides no

account of levels. No answers are given to the key questions about levels outlined above (see section "Basic Distinctions and Defining Questions"). Consequently, we are left in a state of uncertainty regarding what, according to the theory reduction model, constitutes a level; how levels are demarcated; and the principle by which levels are ordered or ranked. It was therefore left to others, namely, Oppenheim and Putnam (1958), to develop an account of levels built upon the traditional logical empiricist framework of explanation and reduction – a conception of levels that remains influential to this day.

Oppenheim–Putnam Levels

Now we are in a suitable position to understand the account of levels that the logical empiricist tradition has inspired, explore its general limitations, and assess its appropriateness as a characterization of the kinds of levels encountered in neuroscience. One lofty goal of logical empiricism was to unify the sciences. Philosophical interest in the possibility of a unified science that would, among other things, provide a general theoretical understanding spanning the full range of natural and social phenomena predates logical empiricism (for further discussion, see Cat 2013; Causey 1977). However, what differentiates the logical empiricists from others with the same objective is that they thought the best strategy to achieve unification involved clarifying the relationships between the various sciences, especially the logical relationships among the theories of those sciences (e.g., Carnap 1955).

Not surprisingly, logical empiricists naturally gravitated towards *reduction* as the relation of choice to characterize the connections between the disciplines and the theories of those disciplines. In "Unity of Science as a Working Hypothesis," Oppenheim and Putnam (1958) present a specific proposal for the unification of science based on successive *micro-reductions* of various higher-level laws or theories (many logical empiricists equate these terms) to a basic or lowest-level physical theory about elementary particles and their interactions. In the course of laying out their reductionistic program, they develop an influential conception of levels subject to considerable discussion in general philosophy of science (e.g., Potochnik and McGill 2012; Rueger and McGivern 2010; Walter and Eronen 2011) and more focused debates about the relation between specific sciences such as psychology and neuroscience (e.g., Bermudéz 2005, 2010; Bickle 2008; Fodor 1974).

Oppenheim and Putnam's (1958) conception of levels is primarily rooted in the existing divisions between the various sciences. Given their general avoidance of metaphysical or ontological concerns, logical empiricists tend not to characterize the natural world itself as comprised of levels of organization. Instead, they identify levels of organization indirectly through the scientific disciplines that investigate them. Along these lines, Oppenheim and Putnam (1958) propose that the different branches or disciplines of science – and the domains of entities associated with the distinctive theories of those branches – each comprise a different level of organization within a single global hierarchy. In total, they identify six levels. Subatomic

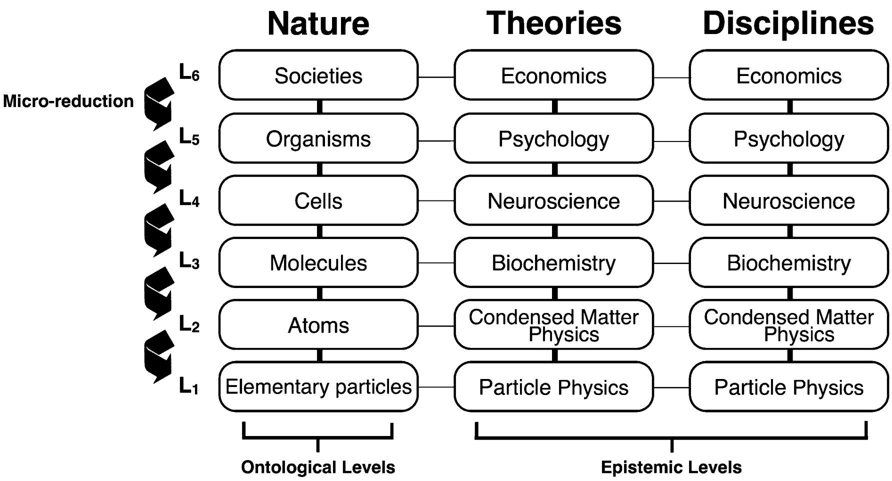

Fig. 2.1 Oppenheim and Putnam's levels hierarchy

physics (also called *particle physics*) occupies the bottom position in the hierarchy as it addresses phenomena at the lowest level of organization, atomic physics (also called *condensed matter physics*) comes next as it targets phenomena at the next highest level, with other disciplines such as biochemistry, biology, psychology, and sociology targeting phenomena at successively higher levels of organization.

For the sake of clarity, these different aspects or dimensions of Oppenheim and Putnam's proposed levels hierarchy – the different disciplines, theories, and domains of objects associated with those theories – can be represented separately (Fig. 2.1). However, it is important to understand that these closely interrelated aspects do not each constitute individual hierarchies in their own right. Oppenheim and Putnam clearly intend to propose only a single, global hierarchy. This more intricate picture helps to illuminate the structure of their account. In particular, it clarifies how the account primarily concerns epistemic levels. Although levels of nature appear to warrant an ontological construal, the fact that this aspect of their levels hierarchy is most accurately described in terms of the domains of objects or ontological posits associated with specific disciplines or theories makes an epistemic interpretation highly plausible (for a similar interpretation, see Craver 2007). Representing things in this way also helps to reveal a costly assumption built into their account, namely, that each organizational level of nature (Fig. 2.1, left column) is associated with one and only one scientific discipline that specifically targets phenomena at that level (Fig. 2.1, middle column) and one and only one corresponding theory specific to that level (Fig. 2.1, right column). The problems associated with this assumption will be discussed in more detail below.

How then does this picture of levels relate back to the goal of scientific unification? Oppenheim and Putnam propose as a "working hypothesis" that the unity of science can be achieved in principle by reducing the theory (or theories) of each higher-level discipline in the proposed hierarchy to those of the next

lower-level discipline in succession until one reaches the lowest-level discipline of fundamental physics dealing with elementary particles. In particular, they argue that this scientific hierarchy supports a special type of theory reduction – *micro-reductions* – which they argue is a satisfactory method for achieving unification. Micro-reductions are a subset of theory reductions in which the objects or entities included in the domain of the reduced theory T_2 are structured wholes that decompose into elements, all of which belong to the domain of the reducing theory T_1 (Oppenheim and Putnam 1958, p. 6). The proposed levels hierarchy is thus ordered so that the scientific discipline (or corresponding theory) addressing entities at a given level can potentially serve as a *micro-reducer* of the discipline (or corresponding theory) addressing entities at the next highest level in the specific sense that the entities from the higher level are decomposable into entities at the lower level. On their account, neuroscience (Fig. 2.1, level 4) serves as a potential micro-reducer of psychology (Fig. 2.1, level 5) – the branch of science at the next highest level in their hierarchy – because the domain of objects countenanced by the discipline of psychology (or by psychological theory) (i.e., organisms) can themselves be decomposed into the entities belonging to the domain of objects countenanced by neuroscientific theory (i.e., cells). The theory at this level can in turn be micro-reduced by the next lowest level, which in turn can be subject to its own micro-reduction, and so on, until some fundamental level is reached.

How well does this account address the aforementioned theoretical desiderata for an account of levels? According to the Oppenheim and Putnam account, the primary relata in the proposed levels relationship are any one or all three of the aspects identified above: object domains, theories, or disciplines. Because they posit a one-to-one-to-one correspondence between levels of nature (domains), theories, and disciplines, these formulations are largely interchangeable. This is their answer to *the relata question*. Next, levels are distinguished and rank ordered by part-whole decomposition. According to their view, any element occupying a higher level is a structured whole that decomposes into parts that are elements in the next lower level of the hierarchy. Cells reside at the next higher level to molecules in virtue of the fact that cells decompose into molecules, molecules are at the next higher level to atoms because molecules decompose into atoms, and so on. Thus, their answer to *the ranking question* is that a given element occupies a different (lower or higher) level than some other element based on a part-whole relationship.

One further feature related to how levels are ranked in the Oppenheim–Putnam hierarchy warrants mention. As indicated above, the levels hierarchy they depict is best understood as a single *global* hierarchy in the sense that every natural system or object must fit into this hierarchy at one and only one particular level. As Bechtel (2008) aptly describes it, the traditional conception assumes that "levels are strata spanning the natural world, with each entity determinately either at the same level or at a higher or lower level than some other entity" (2008, p. 147). It provides a global identification of levels. Finally, according to their account, any whole that is decomposable into parts belonging to a given level will occupy the same level as any other whole that is equivalently decomposable into parts belonging to that same

level (e.g., a pod of Orcas and a colony of ants belong to the same level since they both decompose into individual organisms). Thus, their answer to *the parity question* also clearly depends on part-whole decomposition.

Problems with the Traditional Approach

Oppenheim and Putnam's account has been subject to a considerable amount of criticism. Some have questioned assumptions about the nature of reduction embodied in the account (e.g., Fodor 1974). Others have challenged its descriptive adequacy as an account of reduction in science as it has proved difficult to locate real examples that satisfy the account (e.g., Sklar 1967). Still others have raised issues concerning its oversimplified portrayal of the relationships between the various sciences and the theories and explanandum phenomena of those sciences (e.g., Wimsatt 2007). Here focus will be placed on the specific problems with Oppenheim and Putnam's scheme as an account of levels of organization in neuroscience.

The first major criticism focuses on the fact that Oppenheim and Putnam assume a meaningful demarcation of levels can be based on the divisions between the scientific disciplines, and relatedly, that different sciences exclusively address phenomena at different levels. This overly regimented picture has appeared to many commentators as difficult to maintain (e.g., Bechtel 2008; Craver 2007; Wimsatt 2007). One need only consider the obvious fact that many disciplines, like physics, deal with phenomena at a great diversity of scales ranging from the microscopic (e.g., quarks and neutrinos) and mesoscopic (e.g., turbulent flows) to the macroscopic (e.g., galaxies and supernovas). In such cases, it is difficult to discern a clear sense in which all these phenomena occupy the same level (Bechtel 2008, p. 143). Similarly, biology and neuroscience appear to address phenomena at a range of levels (on most reasonable accounts of levels) (e.g., Churchland and Sejnowski 1992; Craver 2007; Schaffner 1993).

Consider the action potential as a paradigmatic example from neuroscience. The target phenomenon itself arguably resides at the cellular level in the Oppenheim–Putnam hierarchy, since an action potential is defined as a transient change in the neuron's overall membrane potential – a property of the cell as a whole. This much is consistent with their scheme. Yet the current mechanistic understanding of the action potential in neuroscience encompasses multiple levels including molecular and atomic levels (Fig. 2.2). Recall that according to their scheme, each of these levels is proprietary to other disciplines, not neuroscience. Neuroscientists have extensively investigated how ion flow underlying action potentials depends on the behavior of proteins embedded in the cellular membrane acting as voltage-gated ion channels (Hille 2001). These membrane components unambiguously reside at the macromolecular level. Ion channels in turn have their own internal working parts, which have also been subject to intense study in neuroscience, including molecular subunits or domains that support different gating and sensor functions (e.g., Catterall 2000). Finally, other critically important aspects of ion channel function

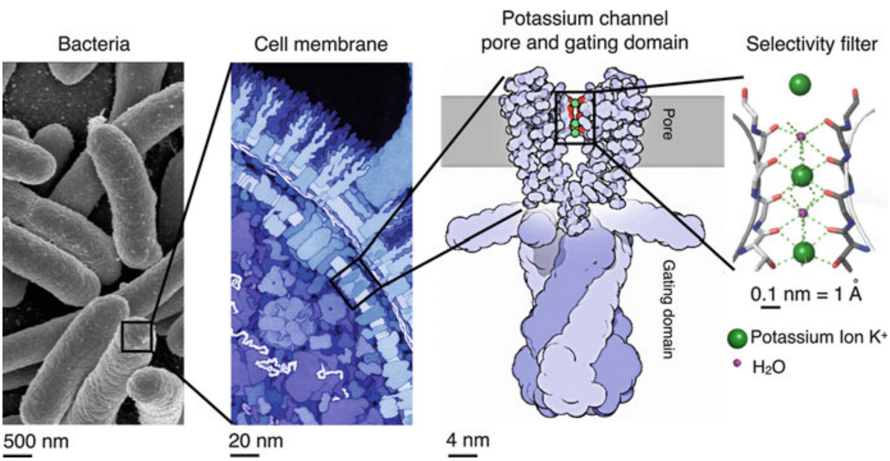

Fig. 2.2 Neuroscience research spans multiple levels of organization (Source: Wikipedia, David S. Goodsell, and RCSB Protein Data Bank. With permission)

such as selectivity for a single ion species are conferred by atomic-scale structural properties of components inside the ion-conducting pore of these channels (e.g., Choe 2002; Doyle et al. 1998). This example of multileveled investigation and explanation in neuroscience is the rule rather than the exception in neuroscience (for additional examples, see Craver 2007). As in the case of physics, it appears deeply misguided to characterize all of these structures as residing at the same level simply because they are addressed by investigators in the same home discipline of neuroscience.

A closely related difficulty stemming from the fact that research in neuroscience typically spans multiple levels is that Oppenheim and Putnam's account is not descriptively adequate since it fails to include the broad range of levels of organization that are vitally important to contemporary neuroscience. They make the strong assumption that there is a one-to-one correspondence between level of phenomena investigated and discipline, but neuroscience clearly does not fit this simplistic scheme. An accurate description of the full range of levels of phenomena addressed in neuroscience would at very least include the circuit level, population level, area level, and network level. In this regard, Oppenheim and Putnam's account is at best incomplete (Craver 2007, p. 173).

A final criticism of Oppenheim and Putnam's account is that it is inconsistent with the fact that different disciplines often target phenomena at the same size or scale (Craver 2007, p. 176). As indicated above, this is especially true for disciplines that investigate phenomena at a wide spectrum of size levels, which notably includes neuroscience. As discussed above, in studying the mechanistic structure of ion channels from the single cell level all the way down to the angstrom scale, neuroscientific investigations overlap in substantial ways with other disciplines including molecular biology, biochemistry, and biophysics in terms of the level of phenomena they target.

Oppenheim and Putnam's account of levels has appeared to many to be fraught with difficulties that are not easily overcome (e.g., Bechtel 2008; Craver 2007). Consequently, many philosophers of neuroscience have chosen to look elsewhere for a conception of levels appropriate to neuroscience.

Explanation and Levels: The Mechanistic Approach

Although the covering law account dominated philosophical thinking about the nature of scientific explanation and reduction for most of the twentieth century, the mechanistic approach is now in ascendancy as a view about explanation in the life sciences including neuroscience. There are two primary reasons for the growing acceptance of the mechanistic approach. First, for many years now, there have been a number of well-known general objections to the covering law account that essentially remain unanswered. These challenges are familiar enough that rehearsing them here is unnecessary (for a comprehensive review, see Salmon 1989). The second reason for abandoning the covering law model centers on the basic fact that laws, and by extension law-based explanations, are few and far between in biological science in general and neuroscience in particular. Instead, there is widespread agreement among philosophers that when investigators in these scientific fields put forward explanations, they often seek to identify the mechanism responsible for maintaining, producing, or underlying the phenomenon of interest. In other words, they seek to develop *mechanistic explanations*. Consequently, over the past two decades or so, philosophers have increasingly focused on the important role that mechanistic explanation plays in the biological sciences including neuroscience (e.g., Bechtel 2008; Bechtel and Richardson 1993/2010; Craver 2007; Machamer et al. 2000). Given its increasing prominence, it should be relatively unsurprising that the mechanistic perspective has also been mined for philosophical insights about the nature of levels.

There are a number of mechanistic accounts in the literature, which differ in subtle and not-so-subtle ways. Nevertheless, a central idea common to all accounts that adequate mechanistic explanations must include descriptions of three key elements: the component parts, the activities of those parts, and the overall organization of those parts and activities that give rise to the behavior of the mechanism as a whole.

Mechanistic Levels

The mechanistic framework embodies a distinctive perspective about levels of organization in mechanisms. The mechanistic view starts with an assumption common to many philosophical theories of levels (e.g., Oppenheim and Putnam 1958; Wimsatt 2007) that the relationship between lower and higher levels is best understood in terms of a compositional or part-whole relationship (Craver 2001;

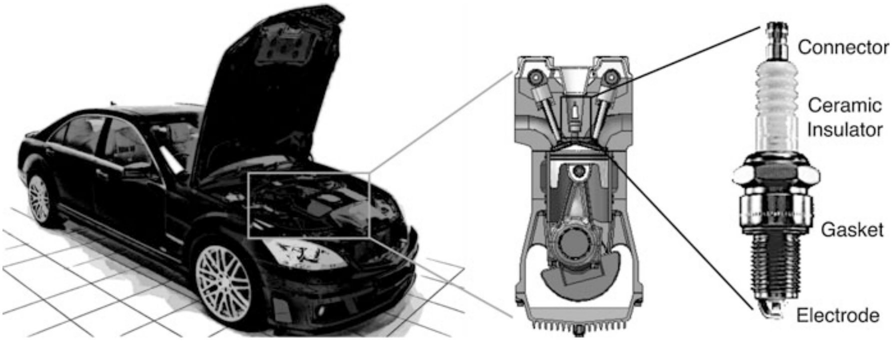

Fig. 2.3 Nested levels of organization in a mechanism

Craver and Bechtel 2007, p. 550). Critically, however, it goes beyond these accounts by imposing an additional requirement that the lower-level parts cannot be any arbitrary parts resulting from any decomposition whatsoever. Instead, the parts occupying the lower level must be working components of the mechanism residing at the higher level.

The mechanistic approach differs from Oppenheim and Putnam's account in one further respect. Since the mechanistic treatment of levels focuses squarely on part-whole relations within the mechanisms themselves to demarcate levels, it is an account of ontological levels (in contradistinction to Oppenheim and Putnam's account of epistemic levels). According to Craver's mechanistic view, "[i]n levels of mechanisms, the relata are behaving mechanisms at higher levels and their components at lower levels" (Craver 2007, p. 189). Bechtel similarly maintains that "[i]t is the set of working parts that are organized and whose operations are coordinated to realize the phenomenon of interest that constitute a level" (Bechtel 2008, p. 146).

Consider how this account of levels applies to a simple mechanism such as the internal combustion engine (Fig. 2.3). The entire working engine system is the overall mechanism responsible for, among other things, producing motion in a car. The engine is composed of a variety of functional parts including intake valves, spark plugs, and pistons, all of which perform activities that contribute in specific ways to the overall performance of the mechanism. According to the mechanistic account of levels, the engine as a whole occupies a higher mechanistic level than that of the component parts, which occupy the next lowest level.

Importantly, mechanistic levels are *nested* in the sense that what comprises a part relative to some higher-level mechanism may itself also be a mechanism relative to some yet lower-level part (Bechtel 2008; Craver 2007). Consider again the internal combustion engine example (Fig. 2.3). While the spark plug is a component in the engine as a whole, and therefore occupies the next lower level, it is also a mechanism in its own right that supports decomposition into its own organized set of parts including the electrode, ceramic insulator, and gasket.

It is also important to recognize that levels of organization in mechanisms are *local* in the sense that they are only defined relative to a given mechanism

(Bechtel 2008; Craver 2007). The mechanistic approach to levels does not propose global strata spanning the natural world as the traditional approach does. In this respect, it is a modest account of levels of organization.

Several important consequences follow from the local identification of levels. First, questions about whether components of a given mechanism (or the mechanism as a whole) reside at a higher, lower, or the same level as entities outside the mechanism are not well defined (Bechtel 2008; Craver 2007). In a particular mechanistic context, two arbitrary elements are deemed to reside at the same mechanistic level only if they are components in the same mechanism. They occupy a higher or lower level depending on how they figure into a componential or part-whole relation within a mechanism. Critically, the mechanistic account provides no resources to compare levels outside a given mechanism.

Second, entities of the same physical type can occupy different levels depending on the specific mechanistic context (Bechtel 2008; Craver 2007). Recruiting the previous example, carbonyl oxygen atoms lining the inner wall of the narrow, approximately 12-Å-long segment of the potassium ion channel pore (Fig. 2.2, rightmost panel) are critical parts of the mechanism for the action potential. These oxygen atoms function as precisely spaced binding sites in the channel's selectivity filter that enable the free passage of potassium ions and impede the movement of sodium ions. As components, they reside one level below the action potential mechanism as a whole. But oxygen atoms are also components of molecules making up many other cellular structures including the cell membrane. As the cell membrane is itself also a component of the action potential mechanism, and so occupies a level below the mechanism as a whole, its own parts will occupy a yet lower level (two levels below the mechanism as a whole). Thus, oxygen atoms reside at different levels depending on how they are employed in the mechanism.

Third, the local character of mechanistic levels allows for a different view of reduction than the traditional theory reduction model (Bechtel 2008). Whereas the traditional approach assumes that higher-level theories can be reduced in succession to increasingly lower levels until some fundamental level is reached, which in turn grounds all the higher levels, the mechanistic approach rejects this global account of reduction. While mechanistic explanations are reductionistic in the sense that they appeal to lower-level parts and their operations to explain some higher-level behavior of the mechanism, they are *locally* reductionistic because there is no single fundamental level that globally grounds all higher levels of mechanisms.

How well does this account address the aforementioned theoretical desiderata? The mechanistic perspective offers clear answers to all three defining questions. First, in contrast to the traditional approach, the mechanistic approach offers an ontologically committed answer to the relata question. The relata are entities that, depending on the particular mechanistic context, play roles either as entire mechanisms or components. Second, the mechanistic answer to the ranking question is that entities are assigned to different levels based exclusively on their part-whole status within a given mechanism. Entire mechanisms occupy higher levels and their components reside at immediately lower levels. As indicated above, the

mechanistic perspective also offers a limited answer to the parity question: entities occupy the same level if they are components in the same mechanism.

Finally, is the mechanistic approach to levels appropriate to neuroscience? Craver has defended this claim in detail (Craver 2007, pp. 165–170), stating that levels of mechanisms best illuminate the sense in which explanations in neuroscience extend across multiple levels. He defends this claim in part by carefully reviewing the multilevel mechanistic explanation of memory formation and spatial learning involving long-term potentiation (LTP) of neuronal synapses in the hippocampus. Instead of walking through that well-known example to show the appropriateness of the account of mechanistic levels, we can again recruit the example discussed above to demonstrate the same thing. Figure 2.2 depicts four distinct levels of mechanisms. There is the level of individual *Escherichia coli* bacteria, the level of the single cell, the level of components of single cells (e.g., ion channels), the level of components of ion channels (e.g., selectivity filters), and the level of components of selectivity filters (carbonyl group oxygen atoms). The specific mechanistic explanation of the action potential makes essential appeal to the last three of these levels. As this and other examples in the literature demonstrate, the mechanistic account of levels provides an accurate characterization of at least one critically important kind of level in neuroscience.

Problems with the Mechanistic Approach

As indicated above, the mechanistic approach to levels offers no means to globally assess level relations beyond the boundaries of a given mechanism. This result will seem especially unsatisfactory to those expecting an account of levels to provide such global assessments and could therefore be viewed as a major deficiency of the account. Proponents of the mechanistic perspective have anticipated this concern and offer two responses. First, they point out that accounts of general or global levels of organization are problematic in their own right. While the project sounds plausible in principle, difficulties emerge when it comes to the actual task of placing all manner of objects into a single levels hierarchy. For example, it is difficult to make sense of the idea of comparing the levels of disparate things like galaxies, elephants, old-growth redwoods, and earthworms. The mechanistic account abandons this project. Second, they maintain that their modest account does the job for which it was intended: it sufficiently captures the sense of levels of mechanisms central to neuroscience research, and this alone should justify its adoption (Bechtel 2008). However, this response merely deflects rather that addresses the worry, because one might continue to hold that any account of levels worth its salt must do more than this. It thus appears that a resolution of this issue demands some prior agreement about what should be expected from an account of levels. Given the state of the dialectic, however, consensus will be difficult to attain.

Another potential shortcoming of the mechanistic account of levels is its treatment of the parity question. Recall that, according to the mechanistic perspective, things occupy the same level only when they are components in the same

mechanism or they stand in a componential relation within a mechanism. This has extremely unintuitive consequences. For example, two objects of the same physical type possessing exactly the same properties will not occupy the same level, if they are not related as components in the same mechanism. Consider the potassium channel example again (Fig. 2.2). According to the mechanistic account, there is no sense in which type-identical H_20 molecules in the cell membrane and in the selectivity filter – both of which are constituents of the same neuron – occupy the same level. This kind of highly counterintuitive result might be taken as evidence that something is deeply problematic with the mechanistic account of levels. Despite these potential limitations, there is much to recommend the mechanistic view of levels as an account of levels in neuroscience.

Levels of Analysis: A Mechanistic Perspective

Another distinct notion of levels that figures prominently in discussions of explanation in neuroscience is the notion of *levels of analysis*. This is the notion of levels at the core of David Marr's (1982) enduring computational framework (For additional discussion of computational issues in neuroscience, see ▶ Chap. 6, "Neural Representation and Computation"). Numerous investigators have since emphasized how understanding the brain requires an approach that successfully combines multiple distinct levels of analysis (Bechtel 1994; Bermudéz 2005, 2010; Carandini 2012; Dawson 1998; Sporns 2010). Nevertheless, some investigators have taken Marr's hierarchical framework to entail an autonomous computational-level explanation of cognitive capacities insensitive to mechanistic details about how such capacities are realized in the brain (e.g., Dror and Gallogly 1999; Rusanen and Lappi 2007). This section briefly sketches how the mechanistic approach affords a fruitful viewpoint on the notion of levels of analysis and, in particular, how it serves to combat a widespread but ultimately mistaken interpretation of Marr's influential framework according to which it supports the idea of an autonomous level of computational explanation. For an expanded discussion of the mechanistic approach to computational explanation in neuroscience, see Kaplan (2011).

Autonomous Computational-Level Explanation?

The idea of an autonomous level of computational explanation has a long history in philosophy and cognitive science (e.g., Fodor 1974; Johnson-Laird 1983; Pylyshyn 1984). According to this perspective, computational explanations of human cognitive capacities can be constructed and confirmed independently of details about how these capacities are implemented in the brain. This claim is also routinely couched in terms of levels. It is often said that the computational-level explanations of psychology are autonomous from implementation-level explanations produced

in neuroscience in the sense that evidence concerning underlying neural organization place no constraints on the shape such higher-level explanations can take.

Time has shown this autonomy claim to be fundamentally mistaken. In particular, scientific efforts to construct autonomous computational accounts unconstrained by data from neuroscience, which could then be mapped back on to real neural mechanisms, are increasingly replaced by psychological models and theories tightly constrained by neuroscientific data. In the absence of constraints derived from what is known about brain structure and function, neuroscientists have been unable to discover the implementations for computational processes posited by high-level psychological theories. This pattern of failure seriously undermines the strict division of intellectual labor assumed by proponents of an autonomous level of computational explanation.

In philosophy, traditional computational functionalism bolstered the idea a *level of autonomous computational explanation*. Functionalists have long maintained that psychology can proceed with its explanatory goals independently of neuroscience. They compare psychological processes to software running on a digital computer. According to the analogy, the brain merely provides the particular hardware on which the cognitive software happens to run, but the same software could be implemented in many other hardware platforms. If the goal is to understand the structure of the software (i.e., what computations are being performed), figuring out the hardware is irrelevant and unnecessary. Despite the fact that it used to be the dominant position in philosophy of mind, functionalism has come under serious challenge. While these lines of argument have run their course, one last arm in the defense of autonomous computational-level explanation remains. It centers on a view about *levels of analysis* famously proposed by the neuroscientist David Marr.

Marr's Levels of Analysis

Marr (1982) distinguishes three levels of analysis intended to apply to information-processing systems ranging from digital computers to the brain: the *computational level* (a specification of the input–output function being computed), the *algorithmic level* (a specification of the representations and computational transformations defined over those representations), and the *implementation level* (a specification of how the other levels are physically realized) (Fig. 2.4).

Unfortunately, several aspects of Marr's own discussion of the relationships between these levels have reinforced the idea of an autonomous level of computational explanation. First, he repeatedly prioritizes the relative importance of the computational level:

> [I]t is the top level, the level of computational theory, which is critically important from the information-processing point of view. The reason for this is that the nature of the computations ... depends more upon the computational problems to be solved than upon the particular hardware in which their solutions are implemented (1982, p. 27).

This privileging of the computational level, coupled with the fact that his preferred methodology is top–down, moving from the computational level to the

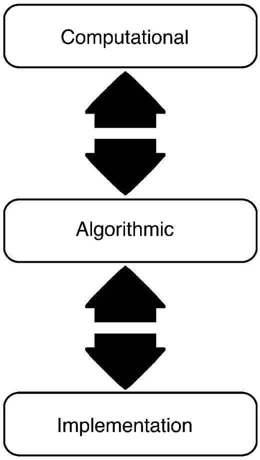

Fig. 2.4 Marr's levels of analysis

algorithmic, and ultimately, to implementation, has fostered the idea of an autonomous level of computational explanation. Second, in some passages, Marr appears to claim that there are either no direct constraints between levels or that the operative constraints are relatively weak and only flow downward from the computational level – claims that are clearly at odds with the mechanistic view. For instance, he states that: "since the three levels are only rather loosely related, some phenomena may be explained at only one or two of them" (1982, p. 25). If computational explanations were unconstrained by one another in this manner, this could be used to draw a conclusion about an explanatorily autonomous level.

When Marr addresses the key explanatory question of whether a given computational model or algorithmic description is appropriate for a specific system under investigation, however, he begins to sound much more mechanistic. On the one hand, his general computational framework clearly emphasizes that the same computation might be performed by any number of algorithms and implemented in any number of diverse hardware systems. And yet, on the other hand, when the focus is on the explanation of a particular cognitive capacity such as human vision, Marr appears to firmly reject the idea that any computationally adequate algorithm (i.e., one that produces the same input–output transformation or computes the same function) is equally good as an explanation of how the computation is performed in that particular system. Consider that after outlining a computational proposal for the extraction of zero-crossings in early vision, Marr and Hildreth immediately turn to a determination of "whether the human visual system implements these algorithms or something close to them" (Marr and Hildreth 1980, p. 205; see also Marr et al. 1979; Marr 1982). This is not an ancillary task. They clearly recognize this to be a critical aspect of their explanatory account. For Marr, sensitivity to information about properties of the human visual system (e.g., the response properties of retinal ganglion neurons) is crucial not only for building successful computer vision algorithms, but also for building successful explanations. According to Marr, the ultimate adequacy of these computational and algorithmic

specifications as explanations of human vision is to be assessed in terms of how well they can be brought into registration with known details from neuroscience about their biological implementation. Marr thus appears to part company with those who maintain that computational explanations are autonomous and unconstrained by evidence about underlying neural systems.

Open questions certainly remain about Marr's computational framework and its relation to explanation in neuroscience that go well beyond the scope of the current chapter. However, this brief exploration of the topic should demonstrate that the mechanistic perspective has great potential to illuminate other kinds of levels in neuroscience besides levels of organization.

Conclusion and Future Directions

There are numerous different kinds of levels in neuroscience that play a variety of important roles. Resources from the mechanistic approach can help to elucidate many of these. It has been argued that the mechanistic approach offers a distinctively useful view about levels of organization in neuroscience. More briefly, it was argued that the mechanistic framework also provides a fruitful perspective on Marrian levels of analysis. A cluster of important questions concerning levels and mechanisms remain open and worthy of further philosophical investigation. One such open question concerns whether global levels of organization can exist in tandem with locally defined mechanistic levels, or whether the two are mutually exclusive, as the current debate implies. This question will be difficult to evaluate until a suitably refined account of global levels is available. Another wide open question concerns how the mechanistic approach handles other kinds of levels including levels of abstraction. Practically everyone acknowledges that scientific explanation – including mechanistic explanation – involves abstraction to some degree, yet how this is compatible with the basic tenets of the mechanistic account remains uncertain. All of these questions about levels are of fundamental importance in philosophy of science and provide worthwhile targets of future exploration.

Cross-References

► Neural Representation and Computation
► Realization, Reduction, and Emergence: How Things Like Minds Relate to Things Like Brains

References

Anderson, P. W. (1972). More is different. *Science, 177*(4047), 393–396.

Batterman, R. W. (2002). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford: Oxford University Press.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines, 4*(1), 1–25.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Lawrence Erlbaum/Taylor & Francis.

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton, NJ: Princeton University Press. Reprinted 2010, Cambridge, MA: MIT Press.

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. New York: Routledge.

Bermúdez, J. L. (2010). *Cognitive science: An introduction to the science of the mind*. Cambridge, UK: Cambridge University Press.

Bickle, J. (2008). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.

Brigandt, I., & Love, A. (2012). Reductionism in biology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2012/entries/reduction-biology/

Bunge, M. (1977). Levels and reduction. *American Journal of Physiology, 233*(3), 75–82.

Callender, C. (1999). Reducing thermodynamics to statistical mechanics: The case of entropy. *The Journal of Philosophy, 96*(7), 348–373.

Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Neuroscience, 15*(4), 507–509.

Carnap, R. (1955). Logical foundations of the unity of science. In O. Neurath, R. Carnap, & C. Morris (Eds.), *International encyclopedia of unified science* (Vol. I, pp. 42–62). Chicago: University of Chicago Press.

Cat, J. (2013). The unity of science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2013/entries/scientific-unity/

Catterall, W. A. (2000). From ionic currents to molecular review mechanisms: The structure and function of voltage-gated sodium channels. *Neuron, 26*(1), 13–25.

Causey, R. L. (1977). *Unity of science*. Dordrecht: Reidel.

Choe, S. (2002). Potassium channel structures. *Nature Reviews. Neuroscience, 3*(2), 115–121.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science, 68*(1), 53–74.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*(4), 547–563.

Dawson, M. R. (1998). *Understanding cognitive science*. Oxford: Blackwell.

Dizadji-Bahmani, F., Frigg, R., & Hartmann, S. (2010). Who's afraid of Nagelian reduction? *Erkenntnis, 73*(3), 393–412.

Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chiat, B. T., & MacKinnon, R. (1998). The structure of the potassium channel: Molecular basis of K+ conduction and selectivity. *Science, 280*(5360), 69–77.

Dror, I. E., & Gallogly, D. P. (1999). Computational analyses in cognitive neuroscience: In defense of biological implausibility. *Psychonomic Bulletin & Review, 6*(2), 173–182.

Findlay, S. D., & Thagard, P. (2012). How parts make up wholes. *Frontiers in Physiology, 3*, 1–10.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175.

Hille, B. (2001) Ion Channels of Excitable Membranes. Sunderland, MA: Sinauer.

Holland, J. H. (2000). *Emergence: From chaos to order*. Oxford: Oxford University Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese, 183*, 339–373.

Lobo, I. (2008). Biological complexity and integrative levels of organization. *Nature Education, 1*(1).

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.

MacMahon, J. A., Phillips, D. L., Robinson, J. V., & Schimpf, D. J. (1978). Levels of biological organization: An organism-centered approach. *Bioscience, 28*(11), 700–704.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 207*(1167), 187–217.

Marr, D., Ullman, S., & Poggio, T. (1979). Bandpass channels, zero-crossings, and early visual information processing. *Journal of the Optical Society of America, 69*(6), 914–916.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace, & World.

Novikoff, A. B. (1945). The concept of integrative levels and biology. *Science, 101*(2618), 209–215.

Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. *Minnesota Studies in the Philosophy of Science, 2*, 3–36.

Potochnik, A., & McGill, B. (2012). The limitations of hierarchical organization. *Philosophy of Science, 79*(1), 120–140.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Rueger, A., & McGivern, P. (2010). Hierarchies and levels of reality. *Synthese, 176*(3), 379–397.

Rusanen, A. M., & Lappi, O. (2007). The limits of mechanistic explanation in neurocognitive sciences. In S. Vosniadou, D. Kayser, & A. Protopapas (Eds.), *Proceedings of the European cognitive science conference*. London: Francis and Taylor.

Sadava, D., Hillis, D. M., Heller, H. C., & Berenbaum, M. (2009). *Life: The science of biology* (Vol. 3). Gordonsville: WH Freeman.

Salmon, W. C. (1989). *Four decades of scientific explanation*. Pittsburgh: University of Pittsburgh Press.

Schaffner, K. F. (1993). *Discovery and explanation in the biomedical sciences*. Chicago: University of Chicago Press.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Sklar, L. (1967). Types of inter-theoretic reduction. *The British Journal for the Philosophy of Science, 18*(2), 109–124.

Sklar, L. (1999). The reduction (?) of thermodynamics to statistical mechanics. *Philosophical Studies, 95*(1), 187–202.

Sporns, O. (2010). *Networks of the brain*. Cambridge, MA: MIT Press.

Walter, S., & Eronen, M. I. (2011). Reductionism, multiple realizability, and levels of reality. In S. French & J. Saatsi (Eds.), *Continuum companion to the philosophy of science* (pp. 138–156). London: Continuum.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Winther, R. G. (2011). Part-whole science. *Synthese, 178*(3), 397–427.

Woodger, J. H. (1929). *Biological principles: A critical study*. London: Routledge & Kegan Paul Ltd.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2009). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2011 ed.). http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/

Experimentation in Cognitive Neuroscience and Cognitive Neurobiology

3

Jacqueline Sullivan

Contents

Abstract

Neuroscience is a laboratory-based science that spans multiple levels of analysis from molecular genetics to behavior. At every level of analysis, experiments are designed in order to answer empirical questions about phenomena of interest. Understanding the nature and structure of experimentation in neuroscience is fundamental for assessing the quality of the evidence produced by such experiments and the kinds of claims that are warranted by the data. This chapter provides a general conceptual framework for thinking about evidence and experimentation in neuroscience with a particular focus on two research areas: cognitive neuroscience and cognitive neurobiology.

J. Sullivan
Department of Philosophy and Rotman Institute of Philosophy, University of Western Ontario, London, ON, Canada
e-mail: jsulli29@uwo.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_108,

Introduction

Neuroscience advances our understanding of the brain and behavior primarily by means of experimentation. Findings from neuroscience shape how we think about the nature of cognition, behavior, diseases and disorders of the mind and brain, consciousness, moral responsibility, and free will. Interpretations of data obtained from neuroscience have the potential to inform diagnostic and treatment decisions and impact assessments of moral culpability and legal responsibility. If the interpretations that neuroscientists make on the basis of data are not warranted, then any claims put forward or decisions made on the basis of that data will lack justification.

Understanding the nature and structure of experimentation in neuroscience and evaluating the explanatory/interpretive claims of neuroscience is crucial for avoiding such epistemological pitfalls. By bringing together insights from the philosophy of neuroscience, the philosophy of scientific experimentation, epistemology and theoretical work in neuroscience and psychology, this chapter puts forward a conceptual framework for thinking about evidence and experimentation in contemporary neuroscience. While the primary focus will be representative examples of experiments undertaken in cognitive neuroscience and cognitive neurobiology, some of the basic lessons are also relevant to experiments conducted in other laboratory-based areas of neuroscience.

Neuroscience and the Experimental Process

One way to think about experimentation in neuroscience and science more generally is as a process, which we may refer to simply as "the experimental process" (See Sullivan 2009). What are the aims of this process? In the simplest terms, the aim of experimentation is to produce data to discriminate among competing hypotheses about a phenomenon of interest. The data from an experiment will serve this function only to the extent that the process of producing those data was *reliable* and the claims made upon the basis of those data are *valid* (See section "Reliability and Validity" below). A worthwhile place to begin to think about evidence and experimentation in cognitive neuroscience and cognitive neurobiology is to say something about the nature of the kinds of claims about phenomena these two areas of science are interested in supporting and the basic types of experiments we find there.

Basic Structure of Experiments: Cognitive Neuroscience

The aims of *cognitive neuroscience* are, roughly, to locate regions of the brain that subserve cognitive functions, to identify patterns of connectivity between different brain regions, and to understand the processing of information through the brain. Cognitive neuroscience combines the conceptual-theoretical framework and

experimental paradigms of cognitive psychology with structural and functional neuroimaging and electrophysiological recording techniques. Experimentation in cognitive neuroscience is based on several basic assumptions. First, organisms have specific kinds of cognitive capacities. Second, these cognitive capacities may be individuated by appropriately designed experimental tasks. Third, for any given cognitive capacity that can be delineated experimentally, it is possible to locate the neural basis of that cognitive capacity in the brain. Identifying the neural basis of a cognitive capacity is assumed to be achievable by correlating (a) subjects' behavioral performance on experimental tasks or their subjective reports with (b) measurable brain activity.

Experiments in cognitive neuroscience combine the use of experimental paradigms/cognitive tasks of cognitive psychology with computational models, neuroimaging, and electrophysiological techniques. A typical experiment in cognitive neuroscience often begins by pointing to previous findings in the empirical literature pertaining to the nature of a specific cognitive function (e.g., face recognition) and the brain area(s) thought to subserve it. These findings are then used as a basis to make testable predictions that are often formulated as competing correlational claims. An example of an empirical question might be: Is the perirhinal (PrC) cortex involved in face recognition? To address this question, an investigator will make predictions about what brain activity in the PrC during a face recognition task ought to look like if it is indeed the case that the PrC is involved. For example, three competing hypotheses may prevent themselves: h_1: Activity in the PrC is increased compared to baseline activity (or activity on a different cognitive task), h_2: Activity in the PrC is decreased compared to baseline activity (or activity on a different cognitive task), and h_3: There is no change in PrC activity compared to baseline activity (or activity on a different cognitive task) (null). In the best-case scenario, the data will adjudicate between these three competing hypotheses and point to the one that is best supported by the data.

While different procedures exist for correlating the behavioral expression of a cognitive function with neural activity (e.g., evoked response potentials (ERPs), positron emission tomography (PET)), by far the most widely employed technology in contemporary cognitive neuroscience, and the one that both philosophers and neuroscientists themselves have questioned the reliability of, is functional magnetic resonance imaging (fMRI). In a typical fMRI experiment, a subject is placed into a magnetic resonance imaging (MRI) scanner and trained in an experimental paradigm. An experimental paradigm is roughly a standard set of procedures for producing, measuring, and detecting a cognitive capacity in the laboratory that specifies how to produce that capacity, identifies the response variables to be measured during pre-training, training, and post-training/testing, and includes instructions on how to measure those response variables using appropriate equipment. It also specifies how to detect a cognitive capacity when it occurs by identifying what the comparative measurements of the selected response variables have to equal in order to ascribe that capacity to a subject (Sullivan 2009). Given that a subject placed in an MRI scanner is physically constrained, the experimental paradigms used in conjunction with fMRI have historically been computer-based

tasks in which the stimuli are presented to subjects on a flat-screen monitor. Subjects elicit behavioral responses to these stimuli or answer questions about them, depending on the instructions provided to them, typically by means of pressing a button.

During task performance or response elicitation, the investigator "scans" the subject's brain, focusing on one or several regions of interest (ROIs). The investigator assumes that when a subject performs a task capable of individuating a discrete cognitive capacity, there will be an increase in neural activity compared to baseline activity in those brain regions involved in task performance. To detect such increases in activity, cognitive neuroscientists rely on the blood-oxygen level dependent (BOLD) response signal. The basic idea is that an increase in neural firing in a given region of the brain triggers a hemodynamic response such that blood is delivered to that area at a more rapid rate than blood that nourishes less active neurons. This increase is accompanied by an increase in oxygen utilization and thus an increase in the amount of deoxygenated blood in the region activated compared to oxygenated blood in the surrounding regions. This difference is used as a contrast for distinguishing areas of heightened activity from areas of less heightened activity. While the subject is in the scanner, sample scans of the brain are taken across a selected time course. Time points of sampling are intended to be coordinated as closely as possible with features of the experimental paradigm such as stimulus presentation and the relevant response output (e.g., button pressing). Once enough data has been collected, investigators pre-process the data in order to eliminate experimental artifacts (e.g., motion of subject while in scanner) and increase signal-to-noise. The data are then processed using statistical analysis techniques (e.g., univariate analysis). The statistically analyzed data is then used as a basis for discriminating among competing functional hypotheses about the brain areas under investigation.

Basic Structure of Experiments: Cognitive Neurobiology

A primary aim of cognitive neurobiology is to discover the cellular and molecular mechanisms of learning and memory (e.g., Sweatt 2009). Cognitive neurobiology combines the behavioral techniques of experimental psychology and electrophysiological, pharmacological, genetic, and protein analysis techniques. A basic set of assumptions informs experimentation in cognitive neurobiology. First, all organisms, from the most simple to the most complex, learn and remember. Second, different forms of learning and memory are detectable by appeal to observable changes in behavior, and may be individuated by appropriately designed experimental learning paradigms. Third, many if not all forms of learning and memory require changes in synaptic strength. Fourth, the changes in synaptic strength that underlie learning and memory are mediated by changes in protein activity and gene expression.

Cognitive neurobiological experiments typically test both correlational and causal or mechanistic claims. Oftentimes an experiment will begin with a question of

whether synaptic, cellular, or molecular activity in the brain *is implicated in* changes in the behavior of an organism or a synapse. For example, in a typical behavioral experiment, an investigator will make predictions about what the measurable changes in cellular and molecular activity and in behavior as a result of training organisms in an experimental paradigm ought to look like in order to establish a correlation between the two factors. Once a correlation between two measurable factors has been established, an investigator typically undertakes intervention experiments. Intervention experiments test predictions about the impact of blocking cellular and molecular activity (with either pharmacological or genetic techniques) on the changes in behavior observed from training organisms in the experimental paradigm. Examples of representative competing hypotheses usually take the following form: h_1: Blocking molecular activity is accompanied by measurable changes in behavior that indicate the learning has been blocked, h_2: Blocking molecular activity is accompanied by measurable changes in behavior that indicate that learning is not blocked, or h_3: Blocking molecular activity results in other measurable changes in behavior that are unexpected (e.g., partial blockade). Again, ideally, the data will adjudicate between the competing hypotheses, discriminating the one that is best supported by the data.

The experimental process in both cognitive neuroscience and cognitive neurobiology is heavily informed and shaped by evidence emanating from other scientific disciplines, including cellular and molecular neuroscience, genetics, psychology, physiology, biochemistry, neuroanatomy, and systems neuroscience. Evidence from these areas serves as a basis for making predictions, formulating testable hypotheses, designing experiments to test those hypotheses, and interpreting the data obtained from such tests. Generally speaking, cognitive neuroscientists and cognitive neurobiologists aim to design experiments to test their predictions and to adjudicate between competing claims about phenomena of interest. When we ask whether neuroscientists succeed at this goal, we are asking whether experimentation in neuroscience is sufficient to yield the data requisite to achieve it.

The Experimental Process

If we want to determine if cognitive neuroscience and cognitive neurobiology are knowledge-generating, it is insufficient to look exclusively at textbook descriptions or reviews of neuroscientific findings. The best unit of analysis is the individual research paper, because it is as close as we can get to the experimental process (without visiting the lab). More specifically, evaluating the merits of already published research papers, which in neuroscience is the aim of lab meetings and journal clubs, is the first step toward answering the question of whether the interpretive or explanatory claims being made in a given research paper are warranted by the data.

Although we often take it for granted that each and every scientist ensures the integrity and reliability of the experimental process himself/herself, it is important to remember that the peer-review process exists in part, because scientists are fallible.

A scientist may believe he/she has adequately tested a hypothesis in instances in which he/she has overlooked potential confounding variables or has failed to exclude or neglected to consider alternative explanations for the results. While peer-review is intended to catch such errors, it offers no full-proof guarantee that science produces knowledge, because individuals on peer-review boards are human and thus fallible, too.

However, even with our unit of analysis being an individual research paper, our access to the experimental process is limited. The introduction provides us with insight into the assumptions that informed a given research study and the origin(s) of the empirical question(s) that the study aims to answer. The methods section provides details about the subjects or objects of the study (e.g., college-age human beings, adult male Wistar rats) and the materials, tools, and techniques used. The results section simply conveys the outcomes of using the methods to answer the empirical question(s). These outcomes are not raw data; they are statistically analyzed data that are typically represented in pictures, charts, and diagrams. The discussion section teases out the implications of the study and attempts to situate the findings within the relevant literature. However, oftentimes the kinds of things that may compromise the knowledge-producing capacity of the experimental process of a given research study are hidden from view. The task is upon us to make the aspects of the process that may potentially compromise the knowledge outcomes of the study (e.g., problematic or unwarranted assumptions, investigator errors, equipment malfunctions, mathematical errors, errors of reasoning and analysis) explicit and to do as thorough a probe of the state-space as possible in order to rule out the possibility that errors were made – to make certain the data can be used to support the interpretative or explanatory claims that the investigators aim to make on the basis of the study.

An appropriate set of analytic tools and a strategy for their application may guide the way. While such tools are not antidotes to error, they at least point us in the direction of where to look for problems in the experimental process that may compromise the ability to use data to substantiate claims about phenomena of interest. What follows is one such set of tools that incorporates insights from philosophy of science, philosophy of experimentation and theoretical work in psychology and the social sciences.

The experimental process has discrete stages (Fig. 3.1). It is set in motion when an investigator or research team poses an empirical question about a phenomenon of interest. Examples of empirical questions in cognitive neuroscience may include: What area(s) of the brain are involved in attention? What kinds of information do dopamine neurons in the ventral striatum encode, represent, or process? What brain areas receive information from mirror neurons? Examples in cognitive neurobiology include: What is the role of protein kinase A in spatial memory? Is activation of cyclic-AMP response element binding protein necessary for long-term potentiation in area CA1 of the hippocampus in vivo? Are synaptic changes that accompany learning and memory similar to those that underlie addiction?

The phrase "phenomenon of interest" is intended to only very loosely capture the idea that *something* prompts an investigator to conduct an experiment – *some*

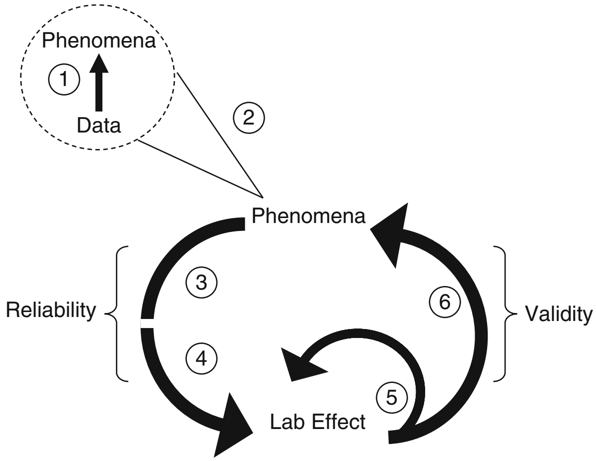

Fig. 3.1 The experimental process. (*1*) An investigator begins with an empirical question about a phenomenon of interest. This question is then redirected at an effect to be produced in the laboratory, thus initiating the (*3*) design and (*4*) implementation stages of data production. If the data production process is reliable, it results in the discrimination of one hypothesis from a set of competing hypotheses about the effect produced in the laboratory. This initiates the stage of data interpretation, in which the discriminated hypothesis is treated as a claim and is taken as true of (*5*) the effect produced in the laboratory and (*6*) the original phenomenon of interest in the world. If the claim was produced by a reliable data production process and it is true of the effect produced in the lab, it is valid (internal validity). If it was produced by a reliable data production process and it is true of the effect in the world, it is valid (external validity) (Sullivan 2009)

phenomenon of interest to him/her. One question that is relevant is how an investigator identifies, detects, or conceives of that phenomenon of interest. An obvious answer, if we look at modern neuroscience, is that in the history of cognitive neuroscience and cognitive neurobiology, some terms have been and continue to be widely deployed, although there is no consensus about how generally to define them. Despite such disagreements, constructs such as attention, working memory, face recognition, and spatial learning, are put forward as starting points for empirical inquiry – investigators pose questions that are directed at shedding light on at least as subset of those phenomena picked out by the concept. One question that remains, though, is how those phenomena that come to be designated by a general construct are identified or detected in the first place. One answer to this question is, if we consider cognitive phenomena more generally, that investigators notice changes in the behavior of organisms from some baseline, which serve as data points for their detection (Step 1 in Fig. 1). Bogen and Woodward (1988), for example, introduce a distinction between "data" and "phenomena" and commit themselves to the idea that phenomena are not observable, but only detectable by means of reference to data, which are observable. However, given that investigators begin experiments with questions about something that *is* detectable, whatever that phenomenon is, it is best understood as detectable derivatively, by means of reference to "data points." For example, most human beings (and non-human

animals) can recognize, after one or more encounters, the faces of conspecifics. This is something that can be detected by noting that on the second or third encounter with a face that was originally novel, an individual's behavior will reflect such recognition. Experiments in cognitive neuroscience and cognitive neurobiology may be said to have their original starting point in such changes in behaviors as exhibited by organisms "in the world." They may also begin with a phenomenon that has been detected by means of data points in the controlled environment of the laboratory (Step 2 in Fig. 3.1). There is most likely a complicated story that could be told as to how an investigator arrived at a particular empirical question. Teasing out this story – looking across review papers and conducting an historical study of the construct/phenomenon in question can be revealing when one attempts to assess the kinds of interpretive claims about a phenomenon that the data obtained from a given research study may be used to support (See, for example, Sullivan 2010).

The experimental process in neuroscience may be regarded as involving two stages: (1) *data production* and (2) *data interpretation* (Woodward 2000). Once an investigator poses an empirical question about a phenomenon of interest, the process of data production begins. Data production may be divided into two discrete stages: (1.1) *design* and (1.2) *implementation*.

The design stage, in basic terms, involves the development of an experimental design and protocol. An experimental design includes the overall set-up of the experiment, in so far as it specifies such things as the experimental context (e.g., how and where objects are to be arranged) and the materials and methods to be used. The experimental protocol is the set of step-by-step instructions that an investigator follows each time he or she runs an experiment. An experimental protocol essentially specifies how each individual experiment is to be run from start to finish. When an investigator is in the middle of an experiment and confused about what to do next – he or she will refer to the experimental protocol (not the experimental design).

Once an investigator has identified a phenomenon of interest, a way to produce that phenomenon of interest in the laboratory must be specified. The phenomenon, whether it is a cognitive function or a form of synaptic plasticity, must be operationally defined. An operational definition is built directly into the design of an experimental paradigm. An experimental paradigm is a standard method or procedure for producing an effect of a specific type. The following features are typically included in the design of experimental paradigms in cognitive neuroscience and cognitive neurobiology: (1) production procedures, namely, a specification of the stimuli (independent or input variables) to be presented, how those stimuli are to be arranged (e.g., spatially, temporally), and how many times they are to be presented during phases of (a) pre-training, (b) training, (c) post-training/testing; (2) measurement procedures that specify the response variables to be measured in the (a) pre-training and (b) post-training/testing phases of the experiment and how to measure them using apparatuses designed for such measurement; (3) detection procedures that specify what the comparative measurements of the response variables from the different phases of the experiment must equal in order to ascribe the cognitive function of interest to the organism, the locus of the function to a given brain area or neuronal population, or a plastic change to a synapse or set

of synapses. This detection procedure is simply an operational definition that specifies the measurable “change” in response variables that must be observed in order to say that the relevant phenomena have occurred.

Investigators in both cognitive neuroscience and cognitive neurobiology have freedom to design experiments – to vary the features of experimental paradigms in ways that they deem most appropriate for their explanatory aims. A “sub-protocol” is a production procedure, written up step-by-step, which corresponds to an experimental learning or electrophysiological stimulation paradigm. It will, for example, specify: (1) the duration of time of the presentation of each stimulus to be used in an experiment, (2) the duration of time that is to elapse between presentation of the stimuli used in an experiment, or the inter-stimulus interval (ISI), (3) the amount of time that is to elapse between individual trials, or the inter-trial interval (ITI), and (4) the amount of time that is to elapse before testing (or biochemical analysis). From the reader’s perspective, the multiplicity of experimental protocols and its implications (Sullivan 2009) are aspects of experimentation that we ought to be privy to when comparing results across different laboratories. This is because subtle changes in experimental paradigms and subprotocols may yield different and sometimes inconsistent results with respect to the phenomenon of interest, rendering it unclear which results should be taken seriously or how to fit the results into a coherent picture or model.

In cognitive neuroscience, the design stage also involves the selection of a subject population, a brain area of interest, experimental techniques (e.g., fMRI, EEG), and statistical analysis procedures (e.g., multivariate pattern analysis (MVPA)). In cognitive neurobiology, the design stage involves the selection of a model organism, a neuronal population or set of synapses, experimental technologies (electrophysiology, biochemistry, immunohistochemistry), and the statistical analysis procedure.

The design stage of data production typically proceeds in discrete stages: Questions are posed and suggestions about how to address them are provided; projections are then made about potential problems that might be encountered in the course of implementing the design and tentative solutions to these problems are offered; and finally, the combined considerations are worked into the design and protocol. Essentially, at this stage, the empirical question of interest is directed at some effect to be produced in the lab.

The implementation stage of data production (Step 4 in Fig. 3.1) begins at some point after an experimental design and protocol has been completed. It involves individual instantiations of the experimental design by means of the systematic following of the experimental protocol using the equipment, materials, and techniques assembled during the design stage. At this point, an investigator takes an individual subject or a group of subjects, and runs them through the steps of the protocol, following those steps as precisely as possible. The immediate output of each individual implementation of the design is an individual data point or set of data points.

Once enough data points for each type of experimental manipulation have been collected, the data points are combined and each complete data set is analyzed statistically. The statistically analyzed data is then used to discriminate one hypothesis from the set of competing hypotheses about the phenomenon of interest

produced in the laboratory. The process of data interpretation then begins. In the first phase of data interpretation (Step 5 in Fig. 1), the hypothesis discriminated by the data is taken as true with respect to the effect produced in the laboratory. That same claim may then be extended back to the original phenomenon of interest in the world that prompted the empirical question in the first place (Step 6 in Fig. 1).

Reliability and Validity

Individual researchers working in laboratories are interested in producing the data requisite to discriminate among competing (correlational, causal, or mechanistic) claims about a single phenomenon of interest. In the ideal case, they aim to design an experiment or set of experiments to produce a set of data *e* in order to adjudicate between a set of competing hypotheses, h_1, h_2, and h_3, about a phenomenon of interest. To do so, the evidence has to be adequate to this purpose. First, the data has to be the outcome of a reliable data production process. What does it mean for a data production process to be reliable? Mayo's (1991) "severity criterion" offers one understanding. In order for a test of a hypothesis to be reliable, it must pass a *severe* test – it must be *highly probable* that the data arising out of a test of a hypothesis would not yield evidence in support of that hypothesis if that hypothesis were in fact false. A related way of understanding reliability is that the process of producing data may be deemed *reliable* if and only if it results in statistically analyzed data that can be used to discriminate one hypothesis from a set of competing hypotheses about an effect produced in the laboratory (See also Bogen and Woodward 1988; Cartwright 1999; Franklin 1986, 1999; Mayo 1991, 1996, 2000; Woodward 1989, 2000). Reliability ought to operate as a constraint on the experimental process. When assessments are made about whether an experiment is reliable given the hypotheses it was designed to discriminate among, how the hypotheses are formulated is fundamental for assessing if the data may serve as adequate evidence.

A second desirable feature of the experimental process, which differs from reliability, is *validity*. Scientific accounts traditionally make use of a general notion of validity, which is taken to be a feature ascribed to experiments or tests. According to these accounts, an experiment is regarded as *valid* if it supports the conclusion that is drawn from its results (e.g., Campbell and Stanley 1963). Scientists and philosophers draw a distinction between external and internal validity (e.g., Cook and Campbell 1979; Guala 2003, 2005). Investigators not only wish to have the conclusions of their results apply to the effects under study in the laboratory (internal validity), they also hope that these conclusions apply to the phenomena of interest at which their empirical questions were originally directed (external validity).

For example, on Francesco Guala's account (2003, 2005), the *internal validity* of an experimental result is established when that result captures a causal relationship that operates in the context of the laboratory. That experimental result is *externally* valid when it captures a causal relationship that operates in "a set of circumstances of interest," outside the laboratory. However, validity may also be understood as a feature of interpretive claims rather than of experimental results. Whereas

experimental results are statistically analyzed sets of data, interpretive claims are what arises when a hypothesis that has been discriminated from a set of competing hypotheses by a set of data is taken as true of an effect produced in the laboratory as well as the original phenomenon of interest outside the laboratory. On this understanding of validity, an interpretive claim about an effect produced in a laboratory, is *internally* valid if and only if that claim is true about the effect produced in the laboratory. A claim about a phenomenon of interest outside the laboratory is *externally* valid if and only if that claim is true about that phenomenon.

One way to understand the relationship between reliability and validity is that they operate as normative constraints on the experimental process, yet give rise to conflicting prescriptions. Reliability prescribes simplifying measures in the context of the laboratory in order to narrow down a set of competing hypotheses about the effect produced in the laboratory. Insofar as it operates to constrain the process of data production, it inevitably restricts the extension of interpretive claims to the laboratory. Validity, however, pulls in the opposite direction. It prescribes that an investigator build into an experimental design those dimensions of complexity that accompany the phenomenon of interest in the world about which an investigator would like to say something. Adhering to the normative prescriptions of validity will inevitably lead to a decrease in the simplicity of the effect produced in the laboratory and an expansion of the set of competing hypotheses that pertain to that effect. In other words, it will lead to a decrease in reliability. However, without reliability, nothing is gained – for if control is lost in the laboratory, nothing true can even be said about the effect produced there – internal validity will be lost as well.

Although not represented explicitly in Fig. 3.1, it is relevant to mention two other types of validity that also may function as constraints on the experimental process. The first constraint is *ecological validity* (See, for example, Bronfenbrenner 1979; Schmuckler 2001). In contrast to external validity, which is concerned with whether an interpretive claim arrived at in a given study may be extended to the real world, ecological validity is concerned with whether the context, stimuli employed, and responses elicited in the experimental context are similar to those that would be found in the world. For example, performing a cognitive task in an fMRI scanner is different than engaging in a cognitive activity in a less restricted environment, so we might say that experiments using fMRI are not ecologically valid. The second type of validity that may constrain the experimental process is *construct validity* (See, for example, Cronbach and Meehl 1955; Shadish et al. 2002). The basic idea here is that investigators in cognitive neuroscience and cognitive neurobiology are interested in developing experimental paradigms that individuate specific cognitive capacities, because they want to be able to make structure-function or mechanistic claims about those capacities. This means that it ought to be the case that the effect under study in the laboratory is an actual instance of the phenomena picked out by a given construct (e.g., "attention"). Notice that if the constraint of construct validity is not met, this poses a problem for reliability – since an investigator may only use data to adjudicate between competing claims about a cognitive capacity produced in the laboratory if it is actually the case that the effect produced by means of an experimental paradigm is an actual instance of that capacity (See Sullivan 2010 for further discussion).

Failure to meet the criterion of construct validity should ideally prompt investigators to look for or develop an experimental paradigm that does a better job at individuating the capacity of interest.

Epistemic Challenges

The conceptual framework offered in the sections above entitled "The Experimental Process" and "Reliability and Validity" may be applied to research papers in cognitive neuroscience and cognitive neurobiology in order to illuminate the steps of the process and identify the various points that decisions are made or courses of action are taken that may impact the reliability of the data production process, the internal and external validity of the interpretive/correlational/causal claims made on the basis of the data, and ecological and construct validity. The framework may also serve as a basis for comparing the experimental process across research studies and determining what kind of interpretive claims are supported by a given body of data. Finally, using this conceptual framework as a backdrop, we can group together epistemic challenges for experimentation in cognitive neuroscience and cognitive neurobiology that have already been identified in the philosophical literature. This is the primary aim of this section.

For example, many philosophers have urged caution with respect to determining the kinds of structure-function claims that fMRI data may be used to support (e.g., Bechtel and Stufflebeam 2001; Bogen 2001, 2002; Delehanty 2007, 2010; Hardcastle and Stewart 2002; Klein 2010a, b; Mole et al. 2007; Roskies 2007, 2010; Uttal 2001, 2011, 2013; van Orden and Paap 1997). A common strategy is to identify the points in the data production process where techniques are used or decisions are made that may jeopardize or compromise the reliability of that process. If we begin by considering the design stage of data production, experiments using fMRI may be regarded as failing with respect to *ecological validity* in so far as subjects perform cognitive tasks that are designed to be implemented while a subject lies still and constrained inside the scanner. However, the cognitive activities that such experiments are supposed to shed light on the neural mechanisms of take place in far less restricted environments. Experiments using fMRI will thus always be limited when it comes to satisfying the criterion of ecological validity. Additionally, given that it is not clear that correlational claims about cognitive functions under study within the confined conditions of the laboratory may be extended beyond that context, fMRI experiments may also be regarded as lacking *external validity*. This does mean, however, that investigators learn nothing about "real-world" cognitive activities when they use fMRI. Rather, it means that we need to think carefully about what kinds of claims are supported by the data.[1]

[1]It is also relevant to note, that some investigators have sought to increase the ecological validity of fMRI experiments with innovative methods that allow for 3-D (as opposed to 2-D) objects to be used within the scanner (See Snow et al. 2011).

A third issue pertains to *construct validity*. For example, performing a face recognition task in a scanner with 2-dimensional stimuli presented on a flat-screen monitor is clearly different from being presented with real 3-D faces. This prompts the question of whether learning about face recognition with 2-D faces is revealing with respect to all of the phenomena that we typically identify as instances of facial recognition, which includes recognition of 3-D faces. A second and related issue, also having to do with construct validity, is whether an experimental paradigm used to study face recognition is sufficient for individuating the cognitive capacity it is intended to measure. For example, face recognition is a complex cognitive function that requires both attentional and mnemonic processes. Thus, an experimental paradigm implemented in an fMRI scanner ought to be able to differentiate the function of attending to faces from face recognition. Although cognitive neuroscientists emphasize the importance of task analysis as a means to ensure the construct validity of their experimental paradigms, they often disagree about which experimental paradigms are the best for measuring different cognitive functions. Such disagreements have prompted skeptics like Uttal (e.g., 2001) to argue that an objective taxonomy of cognitive functions will never be possible. However, some philosophers regard this as far too skeptical a conclusion (See, for example, Landreth and Richardson 2004).

The important point is that if concerns about ecological, external, and construct validity do not shape the development of an experimental design and experimental paradigm and protocol, this will likely impact the kinds of interpretive claims that are warranted by the data. In contrast, since hypotheses typically make reference to the cognitive capacity of interest, if the experimental paradigm is insufficient for individuating that discrete cognitive capacity, then the data will be unreliable for discriminating among competing hypothetical claims pertaining to that cognitive capacity.

Philosophical scrutiny has also been directed at the reliability of data production processes that involve the use of fMRI technology. For example, philosophers have pointed to the fact that the occurrence of the BOLD signal does not directly correlate with task-related neural activity, thus making it a potentially unreliable indicator of such activity (e.g., Bechtel and Stufflebeam 2001; Bogen 2001, 2002; Klein 2010a, b; Roskies 2007). Third, investigators are not always able to distinguish task-related effects from mere artifacts when looking at the raw data. Guesswork is typically required to improve the signal-to-noise ratio in data collected from each subject (i.e., within-subject data) and to eliminate artifacts (e.g., head motion during scanning) before processing the data. Such guesswork leaves open the possibility of experimenter error. Fourth, the fMRI data for each experimental condition has to be determined and averaged across subjects. This requires that the data be mapped and fitted onto an atlas of the brain (e.g., Talaraich atlas). Given differences in the shape and sizes of subjects' brains, averaging the data across subjects and fitting it into the atlas leaves open the possibility of data distortion. Another problem concerns the method of subtraction. In order to determine which area of the brain is involved in which cognitive task, investigators compare the BOLD signal observed on two task conditions that are thought to differ exclusively

with respect to one cognitive activity. For example, face recognition is thought to involve familiarity as well as recollection. One might thus imagine that a subject could be run in a face recognition paradigm and a familiarity paradigm and that activity observed in the familiarity paradigm could be subtracted from that in the face recognition paradigm to yield that area of the brain that is relevant for recognition. However, this method assumes that the two tasks actually discriminate between these two cognitive capacities, which may not be the case (For further discussion, see Bechtel and Stufflebeam (2001); Bogen (2001, 2002); Klein (2010a, b); Roskies (2007, 2010)).

A final issue with fMRI concerns what can be concluded on the basis of fMRI images. As several philosophers have argued, fMRI images are themselves outcomes of data interpretation rather than products of the data production process (e.g., Bogen 2002; Klein 2010b; Roskies 2007). Thus, conclusions that are made on the basis of these images – i.e., using the images themselves to adjudicate among competing hypotheses concerning structure-function relationships in the brain – will fail if decisions made during the stages of data processing involve the introduction of errors that fail to preserve the integrity of the raw data. This is one reason why philosophers have argued that analytic scrutiny must be directed at the analytical techniques involved in the production of fMRI images (e.g., Bogen 2002; Klein 2010b; Roskies 2010).

Despite the apparent limitations of fMRI technology, it continues to be widely used in cognitive neuroscience. Many neuroscientists, however, are aware and openly acknowledge these limitations and are in search of more reliable approaches to locating regions of the brain that subserve cognitive functions, identifying patterns of connectivity between different brain regions, and understanding the processing of information through the brain (See, for example, Logothetis 2008; Culham 2013 http://culhamlab.ssc.uwo.ca/fmri4newbies/).

Cognitive neurobiological experiments have also been a target of philosophical analysis. First, when it comes to the process of data production, cognitive neurobiologists have traditionally been concerned almost exclusively with reliability of the data production process and less concerned with issues of external, ecological, and construct validity. Given that investigators aim to establish causal relationships between cellular and molecular activity and behavior, in order to rule out the possibility of confounding variables, animal subjects are often raised in impoverished environments and trained with types of stimuli having parameters they would be unlikely to encounter in the real world (See Sullivan (2007; 2009) for further discussion). The dissimilarity between the laboratory and the external world thus jeopardizes the ability to extend causal claims established in the laboratory to real-world contexts (See Sullivan (2009) for further discussion).

Another issue that arises with respect to experiments in cognitive neurobiology is that not all investigators are concerned with construct validity. Many investigators are less interested in the cognitive processes that occur when an animal is trained in an experimental learning paradigm than with obtaining data that indicates that an observable change in behavior has occurred. Such data is then used as a basis

for inferring that the cognitive function that the paradigm purportedly individuates has been detected. However, sometimes it is unclear what cognitive capacity a given experimental paradigm actually individuates, which compromises the ability to use data collected using that paradigm as a basis for making causal claims about the role of cellular and molecular activity in a discrete cognitive function (See Sullivan (2010) for further discussion).

Philosophers have also addressed the question of whether results from experiments using model organisms, which are commonplace in low-level neuroscience and the neurobiology of learning and memory, are extendable to the human case (e.g., Ankeny 2001; Burian 1993; Schaffner 2001; Steel 2008; Sullivan 2009). Model organisms include, for example, rodents, sea mollusks, and fruit flies. These organisms are referred to as "models" in so far as scientists use them to establish causal relationships that they aim to generalize to the human population. However, differences between the two populations (i.e., laboratory animals and human beings) and the two contexts (lab versus ordinary environment) complicate the extrapolation of findings from the one context to the other. This prompts the question: When is the extrapolation of causal claims from the one context to the other warranted? Proponents of extrapolation, such as Daniel Steel (2008), have sought to provide an account that puts the investigative strategy on firmer epistemological footing. Of course, strategies for improving extrapolation from model organisms to the human case will vary depending upon the kinds of causal claims at issue (e.g., Sullivan 2009).

Conclusion

Philosophical work on the epistemology of experimentation in neuroscience, as is evident in the above examples, has been directed primarily at the knowledge-generating capacities of specific investigative strategies, tools, and techniques. However, neuroscience is a rapidly expanding field with global aims, the achievement of which requires the development of new and complex technologies. Ideally, we would like to have a workable set of analytic tools that we could apply in different areas of neuroscience and direct at different investigative strategies with the aim of determining whether those investigative strategies are knowledge-generating. Identifying one general set of conceptual tools has been the aim of this article.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Human Brain Research and Ethics
▶ Neuroimaging Neuroethics: Introduction

References

Ankeny, R. (2001). Model organisms as models: Understanding the 'lingua franca' of the human genome project. *Philosophy of Science, 68*, S251–S261.

Bechtel, W., & Stufflebeam, R. S. (2001). Epistemic issues in procuring evidence about the brain: The importance of research instruments and techniques. In W. Bechtel, P. Mandik, J. Mundale, & R. S. Stufflebeam (Eds.), *Philosophy and the neurosciences: A reader* (pp. 55–81). Oxford: Blackwell.

Bogen, J. (2001). Functional imaging evidence: Some epistemic hotspots. In P. K. Machamer, P. McLaughlin, & R. Grush (Eds.), *Theory and method in the neurosciences*. Pittsburgh: University of Pittsburgh Press.

Bogen, J. (2002). Epistemological custard pies from functional brain imaging. *Philosophy of Science, 69*, S59–S71.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*, 303–352.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge: Harvard University Press.

Burian, R. M. (1993). How the choice of experimental organism matters: Epistemological reflections on an aspect of biological practice. *Journal of the History of Biology, 26*, 351–367.

Campbell, D. D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cook, T. D., & Campbell, D. D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Culham, J. (2013). Functional imaging for newbies. (http://culhamlab.ssc.uwo.ca/fmri4newbies/)

Delehanty, M. (2007). Perceiving causation via videomicroscopy. *Philosophy of Science, 74*(5), 996–1006.

Delehanty, M. (2010). Why images? *Medicine Studies, 2*(3), 161–173.

Franklin, A. (1986). *The neglect of experiment*. New York: Cambridge University Press.

Franklin, A. (1999). *Can that be right? Essays on experiment, evidence, and science*. Boston: Kluwer.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science Supplement, 70*, 1195–1205.

Guala, F. (2005). *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Hardcastle, V. G., & Stewart, C. M. (2002). What do brain data really show? *Philosophy of Science, 69*, S72–S82.

Klein, C. (2010a). Philosophical issues in neuroimaging. *Philosophy Compass, 5*(2), 186–198.

Klein, C. (2010b). Images are not the evidence in neuroimaging. *British Journal for the Philosophy of Science, 61*(2), 265–278.

Landreth, A., & Richardson, R. C. (2004). Localization and the new phrenology: A review essay on William Uttal's The New Phrenology. *Philosophical Psychology, 17*, 108–123.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*, 869–878.

Mayo, D. (1991). Novel evidence and severe tests. *Philosophy of Science, 58*, 523–552.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. (2000). Experimental practice and an error statistical account of evidence. *Philosophy of Science, 67*(3), S193–S207.

Mole, C., Plate, J., Waller, R., Dobbs, M., & Nardone, M. (2007). Faces and brains: The limitations of brain scanning in cognitive science. *Philosophical Psychology, 20*(2), 197–207.

Roskies, A. (2007). Are neuroimages like photographs of the brain? *Philosophy of Science, 74*(5), 860–872.
Roskies, A. (2010). Neuroimaging and inferential distance: The perils of pictures. In M. Bunzl, & S. Hansen (Eds.), *Foundations of functional neuroimaging*. Cambridge: MIT Press.
Schaffner, K. (2001). Extrapolation from animal models: Social life, sex, and super models. In P. K. Machamer, P. McLaughlin, & R. Grush (Eds.), *Theory and method in the neurosciences*. Pittsburgh: University of Pittsburgh Press.
Schmuckler, M. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*(4), 419–436.
Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
Snow, J. C., Pettypiece, C. E., McAdam, T. D., McLean, A. D., Stroman, P. W., Goodale, M. A., & Culham, J. C. (2011). Bringing the real world into the fMRI scanner: Repetition effects for pictures versus real objects. *Science Reports* 1, Article 130.
Steel, D. P. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.
Sullivan, J. A. (2007). *Reliability and validity of experiment in the neurobiology of learning and memory*. Dissertation, University of Pittsburgh.
Sullivan, J. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese, 167*, 511–539.
Sullivan, J. (2010). Reconsidering "spatial memory" and the Morris water maze. *Synthese, 177*, 261–283.
Sweatt, J. D. (2009). *Mechanisms of memory*. San Diego: Elsevier.
Uttal, W. R. (2001). *The new phrenology*. Cambridge: MIT Press.
Uttal, W. R. (2011). *Mind and brain: A critical appraisal of cognitive neuroscience*. Cambridge, MA: MIT Press.
Uttal, W. R. (2013). *Reliability in cognitive neuroscience: A meta-meta-analysis*. Cambridge: MIT Press.
Van Orden, G., & Paap, G. C. (1997). Functional neuroimages fail to discover pieces of mind in parts of the brain. *Philosophy of Science, 64*(S1), S85–S94.
Woodward, J. (1989). Data and phenomena. *Synthese, 79*, 393–472.
Woodward, J. (2000). Data, phenomena and reliability. *Philosophy of Science, 67*(3), S163–S179.

Realization, Reduction, and Emergence: How Things Like Minds Relate to Things Like Brains

4

Kenneth Aizawa and Carl Gillett

Contents

Abstract

Debates over reduction and emergence have raged in both philosophy and the sciences. Both sets of debates are surveyed to illustrate the ongoing discussions driven by the sciences and why they are rather different than philosophers commonly suppose. Unlike philosophical debates, scientists on both sides of their discussions accept that composition is universal in nature, that multiple realization exists, and that higher sciences are indispensable. The ongoing debates instead focus on the ontological structure of a certain kind of common case, including what entities are determinative and the varieties of determination.

K. Aizawa (✉)
Rutgers University - Newark, Newark, NJ, USA
e-mail: ken.aizawa@gmail.com

C. Gillett
Northern Illinois University, DeKalb, IL, USA
e-mail: cgillett@niu.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_6,

Introduction

Lively debates rage in the sciences over reduction and emergence. On one side, there are scientific reductionists like Steven Weinberg or E.O. Wilson pressing increasingly sophisticated positions under the slogan that "Wholes are nothing but their parts" (Weinberg 1994, 2001; Wilson 1998). On the other side there are scientific emergentists such as Philip Anderson, Robert Laughlin, or Walter Freeman whose views can be sloganized as "Wholes are more than the sums of their parts" but also that "Parts behave differently in wholes" (Anderson 1972; Freeman 2000; Laughlin 2005). These scientific discussions utilize imprecise understandings of reduction and emergence and also of compositional relations, i.e., "parts" and "wholes," but connect directly to scientific explanations, concrete cases, and hence current empirical evidence.

Reduction and emergence have also been the focus of long-standing debates in philosophy. And there are now also philosophical accounts of scientific relations of composition such as the so-called realization between properties. In fact, on reduction, emergence, and composition, there are an array of competing philosophical frameworks that are very precise, but which are also often technical and inaccessible in character, and framed at some distance from concrete scientific cases and evidence. Although a potential resource, this array of philosophical machinery often makes it hard to see the wood for all the trees.

Given the state of contemporary debates, the primary goal of this survey is to provide the reader with an understanding of the bigger picture that will allow them to see what drives the debates, the nature of the competing positions, and why the issues between them may matter. Previous philosophical discussions of reduction and emergence have simply ignored scientific debates and their positions, but this exclusion looks unprincipled. This survey therefore engages both scientific and philosophical accounts. In fact, the approach here is to start with scientific phenomena and positions, since this reengages the core empirical evidence that plausibly generates the debates. However, the secondary goal of the survey is to provide the reader with an overview of the main philosophical frameworks on each topic.

Overall, the survey shows that the nature of the debates is rather different than philosophers have assumed with regard to both their deeper issues and live positions. Rather than following previous philosophical debates in battling over whether composition is universal in nature, multiple realization exists, or higher sciences are indispensable, for example, all the live positions in ongoing scientific debates over reduction and emergence now accept these claims. Instead, the debates in the sciences focus on disputes about the ontological structure of key scientific cases and disagree over the entities that are determinative, and the varieties of determination, in such examples.

Part 1 surveys the compositional notions posited in the key kind of mechanistic explanation in the sciences and reviews some of the philosophical frameworks used to articulate them. Using this platform, Part 2 highlights how scientifically based forms of *ontological* reduction are live positions that avoid the famous philosophical objections that work against the *semantic* reduction championed by

the Positivists. Lastly, Part 3 distinguishes various notions of emergence, considers whether they challenge such a reductionism in the sciences, and highlights one form of emergentism that does so. Overall, the survey consequently highlights the shape of ongoing debates over reduction and emergence in the sciences.

Part 1: Mechanistic Explanation and the Nature of Compositional Relations

Scientific reductionists and emergentists all focus on what can be termed "inter-level mechanistic explanations" which explain some higher-level entity using lower-level entities taken to compose it. To illustrate the features of these explanations, and their compositional concepts, consider the current scientific account of the growth of dendritic spines in neurons whose basis is familiar from the explanations of cellular motility.

Putting aside all kinds of messenger and other proteins, and a lot of complexities, dendritic spines grow through the polymerization of G-actin (see Fig. 4.1). Multiple G-actin filaments polymerize in certain directions, pressing against other molecules embedded in the cell membrane, thereby reshaping the cell. In this account, scientists postulate a number of compositional relations. There are *part-whole* relations between individuals, since various molecules are taken to be parts of the neuron. There are *realization* relations between property instances, because various properties of molecules result in a property of the neuron. Finally, there is the *implementation* of processes, since the neuron's process of growing a dendritic spine is composed by processes of actin polymerization.

These compositional relations between different categories of entity all share characteristics. Such relations contrast with causal relations, since (i) composition is a species of *noncausal determination relation* that is not identical to the triggering and manifestation of powers, is synchronous, involves entities that are not wholly distinct (i.e., are in some sense the same), and does not involve the transfer of energy and/or mediation of force. (ii) Compositional relations are plausibly *asymmetric*, *irreflexive*, and *transitive*. (iii) Scientific composition also always involves "teams" of individuals, which bear spatiotemporal, productive, and/or powerful relations to each other and are spatially contained within the individual associated with the composed entity. Consequently, (iv) such composition relations are a species of *many-one relation* with teams of many components, whether individuals, property/relation instances, or processes, and one composed entity. (v) Compositional relations have *qualitatively distinct relata* and hence involve "qualitative" emergence, because composed individuals always have different powers and properties, hence grounding different processes, than the powers, properties, and processes found in their components. Thus the cell has the property of growing spines and the powers and processes associated with it, while G-actin has no such properties, powers, or processes. (vi) Scientific composition is plausibly a species of *natural necessitation* where at a time (or across some duration of time in the case of the implementation of processes), the components suffice for the

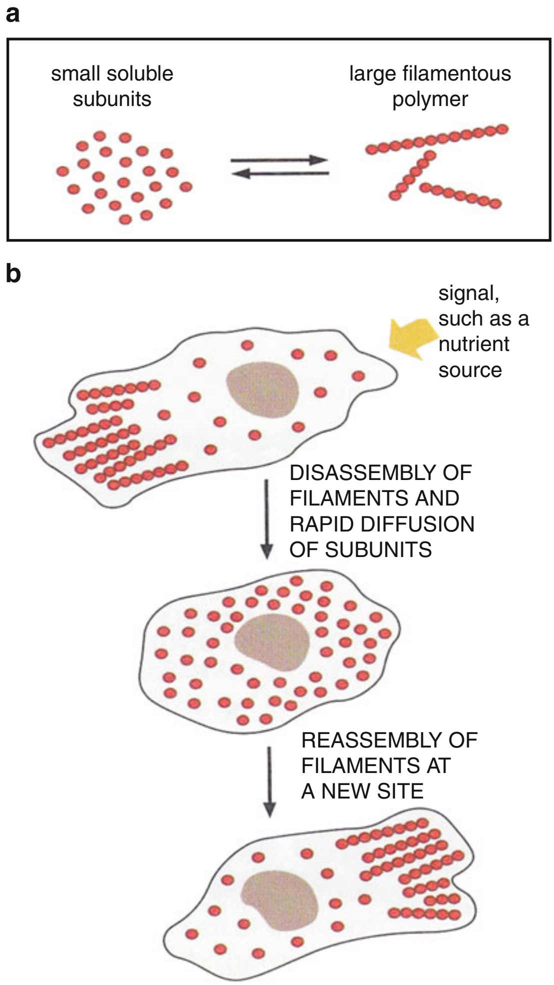

Fig. 4.1 Diagram of the mechanism underlying cell motility and, in the last two steps, the growth of dendritic spines (From Alberts et al (2007), p. 971)

composed entity under the relevant conditions. Thus, when there are certain proteins, with specific relations, under the relevant conditions, there must be the cell. And (vii) scientific composition relations always hold under certain background conditions.

These few characteristics suggest inter-level mechanistic explanations work by representing compositional relations between entities. For, as was just illustrated in (vi), composition is plausibly a species of natural necessitation. Consequently, in an inter-level mechanistic explanation, the lower-level entities of the explanans necessitate, through their compositional relations, that the entities of the explanandum exist under the relevant conditions. Hence the explanans of an inter-level mechanistic explanation plausibly explains its explanandum.

Philosophical attempts to more precisely articulate the nature of scientific notions of composition constitute a very lively area of ongoing research. In the 1980s, philosophers sought to use species of "supervenience" to understand scientific composition. Basically, supervenience is a relation of necessary covariation between entities. The kinds of supervenience thus differ in their relata and the strength of necessity that holds between them. From the 1990s onwards, however, philosophers have rejected supervenience as an account of composition, since supervenience allows covarying entities that are wholly distinct, whereas the relata of compositional relations are in some sense the same (Horgan 1993). [Recall condition (i)].

Many philosophers interested in scientific composition have subsequently shifted their attention from supervenience to so-called realization relations between property instances, so the compositional relations between individuals and processes have been comparatively neglected. Plausibly taking their inspiration from scientific explanations, the two popular kinds of account of realization focus on the differing versions of the idea that components do the causal "work," or as it is often put "play the causal role," of the composed entity.

"Flat" or "subset" views of realization are most popular (Kim 1998; Wilson 1999; Shoemaker 2001). These views take realization to involve causal role-playing through overlap of powers because the realized property instance is individuated by a subset of the powers of the realizer instance. So when one has the powers of the realizer, one must also have the individuative powers of the realized property and hence have a realized instance as well. Realization is thus taken to be a one-one relation of properties of the same individual that overlap in powers and hence overlap in causal role or "work."

One objection to the flat/subset view is that in inter-level mechanistic explanations, such as the example of dendritic spine growth above, realization is apparently a many-one relation between property instances of distinct individuals and has qualitatively different relata that do not share powers (Gillett 2002). These difficulties have prompted alternative, "dimensioned" accounts of realization positing a team of property instances, of component individuals, that realize another qualitatively distinct property instance, of a composed individual, with which they share no powers (Gillett 2002, 2007 and possibly Shoemaker 2007). Here the idea is that many realizer instances can together jointly fill the role of a realized property. Dimensioned views are thus claimed to better fit the features of the relations posited in real scientific explanations.

To summarize, mechanistic explanations often posit compositional relations between the entities found at different levels and studied by different sciences. Such compositional relations are a species of noncausal, synchronous, natural necessitation where, at a time and under the relevant conditions, the component entities suffice for the composed entity because the "work" of the components results in the "work" of the composed entity, whether through role-playing by overlap or joint role-filling. With these points in hand, one can now better appreciate what drives reductionism and emergentism about the sciences.

Part 2: Ontological Versus Semantic Reductionism in Science and Philosophy

It is important to note a common usage of "reductionist" to mean anyone seeking inter-level mechanistic explanations. However, both sides in scientific and philosophical disputes over reduction and emergence now accept the need to search for inter-level mechanistic explanations. So, under the common usage, *everyone* is a "reductionist"! The common usage thus unhelpfully slides over substantive differences between the parties, so it is better to use "everyday reductionism" to refer to the search for inter-level mechanistic explanation.

Scientists like Weinberg or Wilson claim that the results of everyday reductionism lead to more substantive conclusions regarding what may be termed "scientific reductionism." The evolutionary biologist George Williams summarizes the scientific reductionist's key contention when he tells us that the nature of inter-level mechanistic explanation allows us to understand "complex systems entirely in what is known of their components parts and processes" (Williams 1985, p. 1). That is, wherever there are inter-level mechanistic explanations of certain higher-level entities, i.e., certain "wholes," through lower-level component entities, i.e., "parts," then scientific reductionists claims that the composed entities are "nothing more than," "nothing over and above," or "nothing but" such components (which include component individuals and their properties and relations to each other) which should be taken to be the only determinative entities, i.e., the only entities that make a difference to the powers of individuals.

Scientific reductionists thus press ontological parsimony arguments using the compositional notions posited in inter-level mechanistic explanation. There are two hypotheses about the determinative entities underlying such explanations. First, there is the usual hypothesis that there are *both* determinative composed *and* component entities. But the scientific reductionist's rival hypothesis is that there are *only* determinative component entities. For the reductionist notes that when there are compositional relations, then components, under the relevant conditions, suffice for the "work" of the composed entity. Consequently, the reductionist claims that their hypothesis can account for the "work" of both composed and component entities. But, claims the reductionist, that means the hypothesis that there are only determinative component entities is equal in its explanatory adequacy to the hypothesis that there are both determinative component and composed entities. And the parsimony principle tells us that when there are hypotheses that are equally explanatorily adequate, then only the simpler of these hypotheses should be accepted. The scientific reductionist thus concludes that for any successful inter-level mechanistic explanation, it should only be accepted that component entities are determinative. If the Eleatic principle is also endorsed that says the only entities that exist are those that are determinative and make a difference to the powers of individual, then the reductionist will conclude that only the existence of component entities should be accepted.

The result is a form of *ontological reductionism*, for a commitment to both determinative composed and component entities has been reduced to a commitment

only to determinative component entities. And scientific reductionists consequently press connected claims about laws and many argue that, ultimately, the only determinative laws of the universe are the laws of physics focused on the basic component entities (Weinberg 2001, p. 115; and Wilson 1998, p. 55).

However, scientific reductionists also emphasize what Weinberg (1994, p. 52) terms the "compromising" nature of their position, since they accept that higher sciences are in principle indispensable for expressing various truths, including true explanations and laws. For example, Weinberg tells us that:

> The study of statistical mechanics, the behavior of large numbers of particles, ... is a separate science because when you deal with very large numbers of particles, new phenomena emerge ... even if you tried the reductionist approach and plotted out the motion of each molecule in a glass of water using the equations of molecular physics ..., nowhere in the mountain of computer tape you produced would you find the things that interested you about water, things like turbulence, or temperature, or entropy. Each science deals with nature on its own terms because each science finds something in nature that is interesting. (Weinberg 2001, p. 40)

As this passage illustrates, scientific reductionists like Weinberg accept that components form powerful relations to each other bringing into existence new collectives of components, hence new phenomena emerge and so too do new levels understood as scales of such collectives of components. Scientific reductionists like Weinberg thus take higher sciences and their predicates to be indispensable for expressing truths about such collectives of components (Gillett 2007, forthcoming).

Scientific reductionism is consequently an ontological reductionism combined with a species of semantic *anti*-reductionism. Thus in a successful inter-level mechanistic explanation of a neuron using molecules, or of our psychological properties using neuronal or molecular components, the scientific reductionist contends that only the molecules or neurons are determinative or even exist. But the reductionist also accepts that there are many true explanations, or laws, that can only be expressed using the predicates of higher sciences, like "neuron" or "believing," which they take to refer to collectives of such components. Scientific reductionism is thus distinctive by:

(a) Accepting the central role in the sciences of inter-level mechanistic explanations and their compositional concepts
(b) Pressing an ontological form of reduction, driven by parsimony arguments focused on compositional concepts, under which only components are determinative or exist and where the only determinative laws are those holding of components
(c) Accepting the in principle indispensability of higher sciences and hence endorsing semantic anti-reductionism
(d) Embracing a complex macro-world in collectives of components where there is qualitative emergence and "More is different."

Listing these features usefully highlights the contrasts with the Positivists' account of "reduction" that continues to configure philosophical discussions.

Abandoning metaphysics, the Positivists famously needed accounts of explanation and reduction not involving metaphysical concepts like the compositional ones

used in inter-level mechanistic explanations. Instead, the Positivists ultimately fixed upon s*emantic* accounts. For example, the Positivists' view of explanation is the "deductive-nomological" or "DN" model (Hempel 1965) that takes all scientific explanations to have the form of deductive arguments explaining a statement describing the phenomena by deriving it from a law statement serving as one of the premises of such an argument. The Positivist account of explanation thus leaves no room for inter-level mechanistic explanations that often do not utilize derivational relations or laws.

The Positivists' account of reduction, in "semantic" or "Nagelian" reduction (Nagel 1961), builds on the DN account of explanation by focusing on the derivation of the law statements of the theories of higher sciences from the law statements of lower sciences in combination with identity statements relating the predicates of the relevant higher and lower sciences. Under their DN account of explanation, such derivations of higher-level laws constituted an explanation of them and thus semantic reduction putatively "reduced" higher-level laws, and associated theories and sciences, to lower-level laws, theories, and sciences.

Nagelian reduction thus plausibly entails that higher sciences are dispensable, at least in principle, and hence that lower sciences can perform all the significant work of the sciences. In addition, the use in Nagelian reduction of identity statements relating the predicates of different sciences implies that all properties of the higher sciences are coextensive with properties of lower sciences and that there is only one level in that of the micro-world of components. In contrast to scientific reductionism, Nagelian reduction thus:

(a) Uses an account of scientific explanation, in the DN view, that excludes many inter-level mechanistic explanations
(b) Pursues a semantic form of reduction focusing on the reduction of higher-level law statements to lower-level statements
(c) Entails that higher sciences are dispensable and lower sciences are semantically omnipotent, hence endorsing a thorough-going semantic reductionism
(d) Implies, through its inter-level identities, that there are no macro-levels in nature

It can quickly be appreciated why it is important to understand both scientific reductionism and Nagelian reduction, and the differences between them, by looking at the famous philosophical critiques of "reduction."

In the 1970s and 1980s, philosophers of science like Hilary Putnam and Jerry Fodor used examples of inter-level mechanistic explanation to argue that there are few reductions as understood by the Positivist account (Putnam 1967; Fodor 1974). Most famously, the so-called "Multiple Realization Argument" was built around the fact that many scientific properties are "multiply realized" by a variety of distinct lower-level properties and hence not coextensive with such properties, thus blocking inter-level identities and the widespread existence of Nagelian reductions. In addition, the "Predicate Indispensability Argument" argued that, in such cases of multiple realization, the predicates of lower sciences often only serve to express the heterogeneity of lower-level properties, thus missing the true explanations concerning the homogeneities at higher level that can only be expressed using

predicates of higher sciences. Consequently, higher sciences were taken to be indispensable and semantic reduction was challenged in another way.

These famous objections are still widely taken by philosophers to show that reductions are unavailable and that reductionism is a dead view. But the underlying assumption is that reduction must be Nagelian reduction and this can now be seen to be a very dangerous assumption. Neither the Multiple Realization Argument nor the Predicate Indispensability Argument plausibly applies to scientific reductionism that actually accepts their conclusions. For scientific reductionism eschews identities and can use multiple realization to drive its parsimony arguments, thus agreeing with the Multiple Realization Argument's conclusion. And scientific reductionism consequently defends the indispensability of higher sciences, hence accepting the conclusion of the Predicate Indispensability Argument.

Scientific reductionism thus teaches us that semantic and ontological forms of reduction need to be carefully separated. And this highlights the importance of recent philosophical work on ontological forms of reduction that plausibly presses similar lessons. Most prominently, Jaegwon Kim (1998) has formulated an ontological model of reduction in his "functionalization" account, and other versions of ontological reductionism have been offered by philosophers of science such as Alexander Rosenberg (2006) and philosophers of mind like John Heil (2003). In addition, Kim's famous arguments about the "problem of mental causation" are similar in their conclusions to the scientific reductionists' parsimony arguments. [See his "Causal Exclusion Argument" (Kim 1993) and "Supervenience Argument" (Kim 1998).]

There are differences between them, but these philosophical and scientific accounts of ontological reductionism can plausibly inform and strengthen each other. The wider points that consequently come into view are as follows: First, there are substantive forms of ontological reductionism driven by inter-level mechanistic explanations and their compositional concepts such as the realization between properties. And, second, the standard objections to reduction only work against Nagelian reduction and leave such a scientifically based ontological reductionism untouched. Contrary to the received philosophical wisdom, and far from being dead, ontological reductionism thus offers live hypotheses about the structure of cases of inter-level mechanistic explanation in the neurosciences and beyond.

Part 3: The Varieties of Emergentism

Though the famous philosophical objections to "reduction" fail to critically engage a scientifically based reductionism, many self-proclaimed "emergentists" have sought to do so. However, discussions of emergence notoriously involve distinct notions of emergence that are too rarely distinguished. This section therefore starts by outlining four of these concepts and assessing whether any of them poses a challenge to scientific reductionism.

The qualitative emergence that goes along with compositional relations was noted earlier. A qualitatively emergent property is a property of a composed individual that is not had by any of its component individuals. Many writers point

to qualitative emergence as significant evidence against reductionism, but this is a mistake when discussing scientific reductionism. For qualitative emergence accompanies the scientific composition that actually drives the parsimony arguments of scientific reductionism.

Scientific work on complex systems has brought what is termed "weak" emergence to prominence (Bedau 1997). A weakly emergent property is a realized property of a higher-level individual for which the higher-level theories and/or explanations about this property cannot be derived from, or dispensed with in favor of, theories and/or explanations concerning realizer properties. Once again, weak emergence has often been taken to block reduction and this is plausibly true for semantic reduction since the required derivation and dispensability are unavailable. However, weak emergence is compatible with scientific reductionism which requires no derivational relations to run its parsimony argument and accepts the indispensability of higher sciences for expressing many truths about nature.

Rather different is "ontological" emergence, for an ontologically emergent property is instantiated by a higher-level individual but is not realized by other properties (O'Connor 1994). The existence of ontological emergence is often characterized by philosophers as being at the heart of debates over reduction – thus making the central issue the existence of uncomposed entities at higher levels in nature. However, it is quickly seen that this misses what is at issue in scientific debates. It is true that the reductionist's parsimony arguments do not apply to ontologically emergent properties because they have no realizers. However, for similar reasons, ontological emergence also does not block the scientific reductionist's arguments that focus on cases where there are inter-level mechanistic explanations that provide qualitative accounts of the components of the relevant higher-level entities including properties. For ontological emergence consequently does not exist in such cases of successful mechanistic explanation, since all the relevant entities, including properties, are composed. Arguments about the implications of ontological emergence therefore leave untouched the core arguments of scientific reductionism about the implications of mechanistic explanations.

A final notion of emergence has been widely rejected by philosophers and scientific reductionists alike. This is "strong" emergence in a realized property instance that is *both* realized *and* determinative – just the combination of features that the scientific reductionists' parsimony argument, or Kim's reasoning, suggests should never be accepted together. However, scientific emergentists have used real examples to defend a species of strong emergence that appears to offer important challenges to a scientifically based ontological reductionism and its arguments.

For example, system biologists defend this kind of emergence about the molecular components of cells (Boogerd et al. 2005), but it is more appropriate here to focus on an instance from the neurosciences. For example, with regard to "emergence" Walter Freeman claims:

> An elementary example is the self-organization of a neural population by its component neurons. The neuropil in each area of cortex contains millions of neurons interacting by synaptic transmission. The density of action is low, diffuse and widespread. Under the impact of sensory stimulation, by the release from other parts of the brain of neuromodulatory chemicals... all the neurons come together and form a mesoscopic

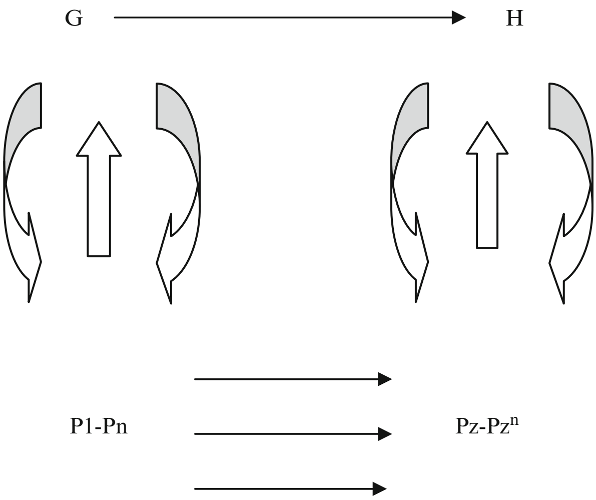

Fig. 4.2 The structure of determinative relations that scientific emergentists claim exist in some cases of inter-level mechanistic explanation. The strongly emergent property instance G is realized (*upward vertical arrow*) by the properties P1-Pn, but G exerts a novel species of determination (*curved downward arrows*) on the powers contributed by these realizers

> pattern of activity. This pattern simultaneously constrains the activities of the neurons that support it. The microscopic activity flows in one direction, upward in the hierarchy, and simultaneously the macroscopic activity flows in the other direction, downward. (Freeman 2000, pp. 131–132)

Let us unpack the interlinked claims in the passage and the position it describes. First, there are no uncomposed entities at the higher level, since the neural population is completely composed. Second, the existence of emergence is not based upon ignorance, but putatively derives from a detailed quantitative understanding of the behavior of components. For, third, writers like Freeman claim that it is now known that components sometimes contribute different powers, and hence behave differently, under the condition of composing a certain higher-level entity where these components would not contribute such powers if the laws applying in simpler aggregations exhausted the laws applying in the complex aggregation. Let us call such powers contributed by a component *differential powers*. Fourth, scientific emergentists like Freeman claim that the differential powers of components are best explained as being determined by the "emergent" entities they compose where this is a novel species of noncausal, but also non-compositional, determination. Lastly, the scientific emergentist claims that new laws about components hold in such circumstances and refer to the composed entity, thus disputing the view of the determinative laws offered by the reductionist.

All manner of issues and questions arise about such claims, but their potential significance for strong emergence is clear. In such circumstances, a realizer property instance, such as a property of a neuron, only contributes certain differential powers when realizing a certain higher scientific property such as one had by a neural population. Scientific emergentists like Freeman take the *realized instance*, the composed entity, to determine the differential powers of their realizers, i.e., their components. At a time, there are thus both the upward and downward determination relations outlined in Fig. 4.2. But when "Parts behave differently in wholes" in this manner,

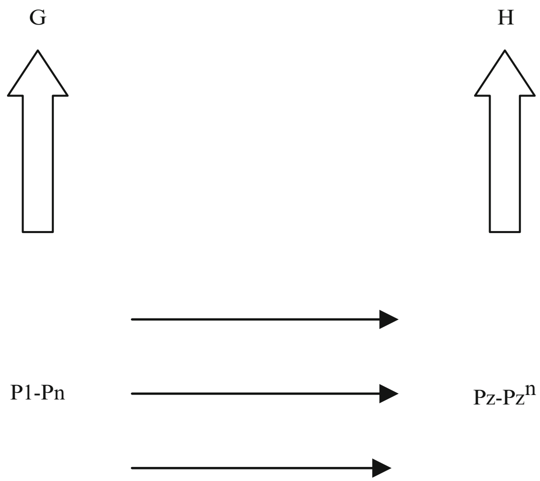

Fig. 4.3 The determinative structure of all cases of composition assumed by scientific and philosophical reductionists. At a time, there are only upward compositional relation between the realizers P1-Pn and the realized property G

then there is an overlooked way, or set of ways, to potentially defend the further claim that "Wholes are more than the sum of their parts." For realized instances, the composed entities, determine some of the powers contributed by their realizers, as well as contributing their own powers, and it appears that there is thus strong emergence in these property instances that are both realized and also determinative.

Unsurprisingly, this kind of scientific emergentism challenges the parsimony arguments of the scientific reductionist or reasoning like Kim's. The scientific emergentist contends that reductionist arguments wrongly assume all cases of composition must have the character outlined in Fig. 4.3 with determinative relations only going upwards at a time. However, if the scientific emergentist's cases are actual, or even possible, then the reductionist reasoning overlooks the different kind of situation outlined in Fig. 4.2. In such cases of strong emergence, it is true that there are compositional relations going upwards, but also true that composed entities are downwardly determinative as well. If its claims can be sustained, scientific emergentism would thus challenge the *validity* of the scientific reductionist's parsimony arguments.

Philosophers, sometimes working together with scientists, have offered theoretical frameworks seeking to articulate this kind of scientific emergentism and its implications (Boogerd et al. 2005; Gillett forthcoming; Mitchell 2009). The resulting forms of scientific emergentism plausibly offer live hypotheses about successful cases of inter-level mechanistic explanation in the sciences that are important competitors to those of scientific reductionism.

Conclusion

This survey illustrates the substantive, ongoing debates in the sciences and philosophy over reduction and emergence. Earlier philosophical debates fought over the

ubiquity of mechanistic explanation or whether compositional concepts apply to all entities and over whether multiple realization exists or higher sciences are dispensable. However, all the live positions in the present debates in the sciences now accept such claims. Scientific emergentism endorses the prevalence of mechanistic explanations and takes all higher-level entities to be composed. While scientific reductionism accepts that higher sciences are indispensable and that multiple realization exists. Ongoing debates about reduction and emergence, apparently tracking empirical evidence, have thus moved on from whether mechanistic explanations *exist* for all the entities in nature to focus on the *implications* of such explanations.

On one side, proponents of a scientifically based ontological reductionism claim that components like neurons, or even molecular constituents, are the only determinative entities in the cases of inter-level mechanistic explanation common throughout the neurosciences and that the productive interactions of such components are the only species of determination in such examples. While on the other side, scientific emergentists contend that components such as neurons, or even molecular entities, sometimes have differential powers and that emergent composed entities, whether populations of neurons, brains, or even psychological entities, downwardly determine that component entities have such powers – hence defending more determinative entities, and species of determination, than the reductionist.

The opposing scientific reductionist and emergentist views thus offer substantively different accounts of the structure of concrete scientific examples in the neurosciences and other sciences. And how the debates over these views are resolved will plausibly decide whether composed entities, like neurons, brains, or even ourselves, are taken to play a determinative role in the universe or whether the scientific reductionist's simpler picture of nature should be accepted.

Cross-References

▶ Consciousness and Agency
▶ Determinism and Its Relevance to the Free-Will Question
▶ Explanation and Levels in Cognitive Neuroscience
▶ Free Will, Agency, and the Cultural, Reflexive Brain
▶ Mental Causation
▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System

References

Alberts, B., Johnson, A., Lewis, J., & Raff, M. (2007). *The molecular biology of the cell* (7th ed.). New York: Garland.

Anderson, P. (1972). More is different. *Science, 177*, 393–396.

Bedau, M. (1997). Weak emergence. *Philosophical Perspective, 11*, 375–399.

Boogerd, F., Bruggeman, F., Richardson, R., Stephan, A., & Westerhoff, H. (2005). Emergence and its place in nature. *Synthese, 145*, 131–164.

Fodor, J. (1974). Special sciences: Or the disunity of science as a working hypothesis. *Synthese, 28*, 97–115.

Freeman, W. (2000). *How brains make up their minds*. New York: Columbia University Press.

Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis, 62*, 316–323.

Gillett, C. (2007). Understanding the new reductionism. *The Journal of Philosophy, 104*, 193–216.

Gillett, C. (forthcoming). *Reduction and emergence in the sciences and philosophy*. Cambridge: Cambridge University Press.

Heil, J. (2003). *From an ontological point of view*. New York: Oxford University Press.

Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Horgan, T. (1993). From supervenience to superdupervenience. *Mind, 102*, 555–586.

Kim, J. (1993). *Supervenience and mind*. Cambridge: Cambridge University Press.

Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.

Laughlin, R. (2005). *A different universe: Reinventing physics from the bottom down*. New York: Basic Books.

Mitchell, S. (2009). *Unsimple truths: Science, complexity and policy*. Chicago: University of Chicago Press.

Nagel, E. (1961). *The structure of science*. New York: Harcourt Brace.

O'Connor, T. (1994). Emergent properties. *American Philosophical Quarterly, 31*, 91–104.

Putnam, H. (1967). Psychological predicates. In W. Capitan & D. Merrill (Eds.), *Art, mind and religion*. Pittsburgh: Pittsburgh University Press.

Rosenberg, A. (2006). *Darwinian reductionism*. Chicago: Chicago University Press.

Shoemaker, S. (2001). Realization and mental causation. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents*. New York: Cambridge University Press.

Shoemaker, S. (2007). *Physical realization*. Oxford: Oxford University Press.

Weinberg, S. (1994). *Dreams of a final theory*. New York: Random House.

Weinberg, S. (2001). *Facing up: Science and its cultural adversaries*. Cambridge, MA: Harvard University Press.

Williams, G. C. (1985). A defense of reductionism in evolutionary biology. In R. Dawkins & M. Ridley (Eds.), *Oxford surveys in evolutionary biology* (Vol. 2). New York: Oxford University Press.

Wilson, E. (1998). *Consilience: The unity of knowledge*. New York: Knopf.

Wilson, J. (1999). How super-duper does a physicalist supervenience need to be? *Philosophical Quarterly, 49*, 33–52.

Mental Causation

5

Holly Andersen

Contents

Abstract

The problem of mental causation in contemporary philosophy of mind concerns the possibility of holding two different views that are in apparent tension. The first is physicalism, the view that there is nothing more to the world than the physical. The second is that the mental has genuine causal efficacy in a way that does not reduce to pure physical particle-bumping. This chapter provides a historical background to this question, with focus on Davidson's anomalous monism and Kim's causal exclusion problem. Reponses to casual exclusion are categorized in terms of six different argumentative strategies. In conclusion, caution is advised regarding the inclination to reduce the mental to the physical and a positive direction is sketched for substantively characterizing mental causation by recourse to well-confirmed accounts of causation coupled with empirical research.

H. Andersen
Philosophy, Simon Fraser University, Burnaby, BC, Canada
e-mail: handerse@sfu.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_110,

Introduction

Cognitive neuroscience is often taken to either imply, or minimally to be compatible with, a view about the nature of mind called physicalism. Physicalism is the view that everything in the world, including but not limited to the mind, is purely physical in character; it essentially denies the existence of nonphysical entities or processes. The debate about the existence and nature of mental causation stems from a tension between a commitment to physicalism concerning the nature of the mind, and the apparent causal efficacy of such mental events or states such as having an intention or committing an action. The commitment to physicalism is inconsistent with "spooky" mental causes, such that thoughts, beliefs, or intentions are not themselves physical, but somehow reach out and poke the physical world. Physicalism is generally taken to imply that, whatever mindedness turns out to be, it should fit clearly into the nexus of physical causes with which we are already familiar. On one hand, this commitment appears very much in line with the approach to studying the mind in cognitive neuroscience. On the other hand, the realm of the mental appears at least prima facie to be genuinely causally efficacious – we do at least *seem* to bring about our actions by deliberation and intentions on which we act – in ways that cannot be straightforwardly reduced to familiar physical causation. This perceived conflict between physical causation and the place of the mental is the source of debate about the existence and nature of mental causation.

The fact that the mental is characteristically ordered by rational norms that do not appear to exist in physical causation pulls toward causal autonomy of the mental from the physical, while the commitment to physicalism resists this as spooky or mysterious. Solving the problem of mental causation requires finding a way to reconcile physicalism about the mind with causal efficacy of the mental, either by situating mental causes in the physical world, by rejecting mental causes, or by rejecting physicalism. This chapter will focus on the first two of these options, since the commitment to physicalism is both widespread in this discussion, and a unifying premise shared by many disparate views.

The problem of mental causation is closely related to, but not the same as, what is often called the mind-body problem. The mind-body problem concerns the relationship between the mind and the body: Is the mind nothing more than the physical body with a certain arrangement of parts? Is the mind a collection of causal functions performed by the body? How does consciousness arise from the body? The problem of mental causation is intricately connected in that many answers to the question of how the mind and body are related will have starkly differing consequences for whether or not the mind has any genuine causal efficacy on the body, or on the world via the body. They are, however, different issues. The question of phenomenal consciousness and its relationship to various neurophysiological processes may be answered, for instance, without thereby yielding a firm answer to the question of whether phenomenal consciousness has genuine causal efficacy, on what it can act, etc.

Historical Background

The original version of the problem of mental causation as it figures in contemporary debates is often taken to begin with Descartes' Meditations (1641/1996), although it can be dated back as far as Plato's *Phaedo*. Descartes presented a dualist picture of the world as comprised of two distinct substances, physical and mental. Physical substances, or objects made of matter, are spatially extended and located, and are not capable of thought. Anything composed solely of matter moves via mechanical forces only; Descartes thought that animals, for instance, were simply very complicated automata. Mental substances, or minds or souls, are that which thinks; they are not spatially extended or located. Properties of physical substances are, on Descartes' view, fundamentally incompatible with mental substances. Minds cannot have shapes or motion, and material objects cannot have thoughts or sensations.

Humans are, according to Descartes, an intimate union of a material body that moves like a machine with a mental substance that thinks, that senses via the material body, and that can influence the motion of that material body. Descartes was careful to avoid a view in which the mind and body were overly separate. The mind-body union is not like the ship in which a sailor directs the motion, he says. A sailor can only know that there is damage to the ship's hull by going and inspecting it. But we can know things about our body in a direct way, by sensing, in a way that we cannot with any other material body to which our thinking minds are not intimately connected. We do not control our body in the distant way in which we control a puppet; there is an immediacy to how we move our limbs that is the consequence of our minds being connected in this special way to our body and not to other pieces of matter.

Descartes thus denies physicalism about the mind-body relationship: The mind could not be made out of particular bits of matter; no amount or organization of matter could ever comprise a mind, because it would always be of the wrong kind of substance to do so. But the mind is intimately connected to some particular chunk of matter, to its body. It receives sensations from that body and, most relevant for our purposes, it also moves that body directly, but no other bodies directly. There are then two directions of influence that pose a problem: How does the physical world have a causal effect on the mind via perception? And, how does the mind influence the body?

Descartes' account is illuminating of the trajectory of discourse on mental causation not only because he runs into a certain kind of problem, but also because he then proposes a certain kind of solution that runs into another very characteristic problem. His dualist account generates a clear case of a problem for mental causation by conceiving of the mind as an entirely different kind of thing than the physical body on which it acts. If physical motion is entirely mechanical, and the mind is entirely unphysical, how does the mind communicate motion to the body? His solution is to propose a way-station where physical bumpings and pullings are translated into signals that influence the mind. Sensation involves physical causation that is transmitted via a complicated series of physical machinery to a special

place in the brain – famously, he proposed the pineal gland as a possible site for this – that translates the physical signal into a mental one. Once the mind resolves to do something, such as lift an arm, that mental signal is then translated back via the same site to a physical signal that would pull and push on the right parts of the body to make the arm go up.

This solution to how the mind and body causally influence one another does not really solve the problem, though. Instead, it relocates it. Proposing a way-station that translates mental and physical influence back and forth between the two kinds of substance physically isolates the location whereby the mind controls the body, but still does not answer the problem of *how*, exactly, the pineal gland translates between two fundamentally different substances. Localizing the physical space in which such a translation happens is arguably an improvement over allowing it to happen all over the body, but still it does not budge the problem of mental causation: How does that mental influence get to the body, and use that body to influence the world? This has been called by Robb and Heil (2013) a problem concerning the causal nexus by which mind and body are connected: "Any causal relation requires a *nexus*, some interface by means of which cause and effect are connected."

The way in which this problem arises in Descartes' account is particular to his dualist views, and the problem of how the mind could exercise a causal influence in the physical world looks somewhat different when one rejects substance dualism. But it is a problem for causation that will arise in a different form anytime the mental and physical are treated as of different sorts: *How*, not merely *where*, does one causally influence the other?

This problem of how the mental could influence the physical simply does not arise in accounts of the mind that treat it as identical with, or at least of the same kind of metaphysical substance as, the brain – in other words, in any physicalist account of mind. In the twentieth century, for instance, the identity theory held that the mind just is the brain – processes in the mind are not merely correlated with, but simply are identical to, processes in the brain (for instance, Smart 1959). In this account, there is no causal influence of the mental beyond that of the causal influence exerted by the brain. Type identity theory is the view that types of mental processes, like pain, simply are types of physical processes, like C-fibers firing. The view that types of mental processes or events are nothing other than types of physical events or processes is also called reductive physicalism: physicalism, because it holds that the mind is physical in character, and reductive, because it holds that the mental reduces to the physical.

While there are different versions of it, reductive physicalism, in any form, solves the tension between a physicalist view of mind and the apparent causal efficacy of the mental by denying that mental causal efficacy is anything other than the causal efficacy of neurophysiological processes. The reductive element of the view means that mental causation is a placeholder for the real causal story which inevitably involves brain processes, microphysical particle states, or some other straightforwardly physical cause. Reductive physicalism about the mind follows a trend in the twentieth century that characterizes many scientifically informed philosophical accounts. It organizes the relationship between types like the mental

and physical in terms of levels, such that the causal efficacy (and, indeed, other features like metaphysical fundamentality) of higher levels depends on and reduces to that of lower levels. The mental is causally efficacious only insofar as it reduces to the physical, and the physical is causally efficacious. Reductive physicalists allow that mental terms are a pragmatic convenience and do not need to be eliminated, so long as they are understood as a kind of short-hand for the "real" causal story. This means that reductive physicalism is less strong than eliminative physicalism, which holds that mental terms will, or at least should, be replaced entirely (P.S. Churchland 1986; P.M. Churchland 1981).

Anomalous Monism and Causal Exclusion

The tension between physicalism and the causal efficacy of the mental was given an influential analysis by Donald Davidson (1980/2001), (1995). He aimed to both preserve physicalism as a view about the nature of the mind as part of the physical world studied by science, while also arguing for autonomy from physical causation for causes and causal relations that involve mental terms. The mental was part of the physical world, but could not be reduced to the laws of physics. His view, anomalous monism, was pitched at least partially as an alternative to reductive physicalism. His account offers a nonreductive physicalist way to accommodate both of the intuitions that there is nothing nonmaterial about mental events and that mental events do have genuine causal efficacy.

This section explores Davidson's account and Kim's challenge to it, and how this sets the framework for much of the contemporary debate about mental causation, as well as connecting that debate to the broader issue of higher versus lower level causation generally.

Davidson subscribes to what he calls the Cause Law thesis (1995), which is one of the key premises in his account and closely related to the deductive-nomological model of explanation in science. According to the Cause Law thesis, all true causal statements are such that some general causal law connects the cause and the effect under some description. Singular causal claims are made true, because there is an underlying general law of the form, All Xs cause Ys, such that the cause and the effect are instances of Xs and Ys. The singular claim, "that rock just broke this window," might be true, because "All rocks, thrown sufficiently hard, break windows," is true. This is where the nomological force of causal relationships comes from – the cause necessitates the effect because of the law connecting them, and without such a law, there would be no connection between cause and effect.

According to Davidson, causes and effects are those things out there in the world, to which we point with our descriptions. As such, Davidson holds that the same causes and effects can be picked out using different descriptions. When we make a true causal claim that x caused y, it is true because of that covering law, regardless of whether the covering law involves x and y so described or if it involves x and y under an entirely different description. Thus, the Cause Law thesis is an existence claim about there being a law under *some* description of the cause

and effect, not that there is a law for every description, nor that there is a law that we know, or even that we know the description under which the law holds. It is merely the claim that, whether or not we ever know it, there is such a description and such a law, and it is this which makes true any causal claim, including but not limited to those involving mental causes.

Putting these pieces together, Davidson holds that mental causes (and effects, although mental causation is primarily taken to be problematic for mental causes of physical events) are, extensionally, part of the same physical causal nexus that is studied by the sciences, and are thus causal, because they figure in true general laws. However, there are no laws containing mental terms – no laws in which both X and Y are mental, nor laws in which either X or Y are mental. Mental causes can be genuinely causal, but mental causes and effects are anomalous, because it is never as mental that they are covered by a law. Mental descriptions are one way of describing the causes and effects in questions, but not the only way. If one were to redescribe those same causes and effects in other terms, namely, in the right physical terms, then a general law of the form All Xs are Ys would cover every instance of true mental causation. But it would never cover it under the mental description.

In this way, Davidson hopes to salvage features of the mental, like rational constraints on action, while still situating mental causation within the same causal nexus as the causes and effects studies in the sciences. The kinds of causal relationships we find in physics cannot account for or accommodate what he called "the uneliminably normative or rational aspect of intentional idioms, and the consequent irreducibility of mental concepts to concepts amenable to inclusion in a closed system of laws" (Davidson 1995). Because the mental is anomalous, with no genuine laws involving mental terms, it cannot be reduced to the merely physical. Mental descriptions continue to be useful as a separate domain of discourse, even while acknowledging that the mental cause and effects in question are not anything above and beyond the physical causal nexus with which we are scientifically comfortable. The mental is simply another way, a nonreducible way, of describing certain parts of that nexus.

There are many issues that could be raised (and have been, by a wide variety of authors) with respect to Davidson's account. For instance, it relies on the rather antiquated deductive-nomological account of explanation, where causal explanations must all be made true by universal laws. Given other accounts of explanation and/or causation, it is not clear that anomalous monism could even be formulated, or that mental causation actually poses any kind of particular problem – we will explore this line of thought more in a subsequent section. Furthermore, it is something of a matter of faith that there really are multiple legitimate descriptions of the *same* mental causal relata. As we will also see in a subsequent section, there are many cases where redescribing causal relata actually changes the subject, by changing the causal relationships into which it enters. But for now, it suffices to lay out Davidson's view of anomalous monism as a key position that aimed to accommodate mental causation within the commitments of physicalism, against which the main challenge against genuine mental causation, that of causal exclusion, is targeted.

Davidson's account is the primary target for Jaegwon Kim's (1989, 2000) causal exclusion argument, which is arguably the core of the contemporary problem of mental causation. Kim challenges the idea that Davidson has salvaged genuine causal efficacy for the mental, and argues that it is instead a different form of reductive physicalism. Kim's challenge to anomalous monism can be generalized to pose an issue not just for mental causes, but for any causes that are in some way higher level with respect to another set of potential causal relata.

Kim's influential work involves both a specific criticism of Davidson's account as well as a broader challenge to any nonreductive account of the relationship between the mental and the physical. One of his main criticisms of Davidson's account of anomalous monism is that it does not actually establish genuine causal efficacy for the mental, and, thus, does not block the reduction of the mental to the physical. That some event has a mental description is causally irrelevant, claims Kim (1989). If it is only under a physical description that an event can be covered by a law, and it is only because of a covering law that an event is causally efficacious, then no mental event is ever causally efficacious because it is mental, but only because it is physical.

Horgan referred to this as the problem of quausation (1989): Mental events enter into causal relationships, but only as physical events, never by dint of their being mental events. The mental is an epiphenomenal ride-along of the physical, meaning that mental causation is merely apparent and nothing other than microphysical causation under a different name. Many authors find this an acceptable conclusion. But, in order to retain physicalism *and* genuine causal efficacy of the mental, something more is needed than simply having any kind of causal efficacy. It needs to be the right sort of causal efficacy, the sort attributable to mentality, not merely physicality.

The broader challenge that Kim has issued is often referred to as the causal exclusion problem. It applies to any nonreductive yet physicalist account of the mental, and sets out what must be overcome in order for mental causation to be genuinely mental, rather than an epiphenomenal redescription of genuine microphysical causation. Kim highlights the tension between physicalism and genuine mental causal efficacy by pointing out that the class of physical causes is supposed to be complete: For any physical effect, there is a sufficient cause of it that is physical, also. This is the rejection of dualism, of "spooky" nonphysical causation somehow bumping parts of the physical world around. According to anomalous monism, some physical causes are also mental causes. However, the effects of those causes are already "caused," as it were: For any physical effect, there is already a complete physical causal story.

One is then confronted with the following dilemma: Either such effects, with both physical and mental causes, are overdetermined or they are not overdetermined. If they are not overdetermined, then there is something that the mental contributes causally to the effect. But this also violates the assumption of physicalism. On the other horn of the dilemma, however, all effects with mental causes are overdetermined, which means that they have multiple causes, each of which is sufficient to bring about the effect. The physical causes alone, with no additional

mental cause, would have been sufficient to bring about the effect exactly as it occurred. This horn has two unfortunate consequences. It renders mental causes systematically superfluous, while also committing to a metaphysically suspect view about the incredibly rampant presence of overdetermination in the world for just a particular set of causes, namely, all and only the mental ones. Neither horn of this dilemma accomplishes the goal of preserving both physicalism and genuine causal efficacy of the mental.

The argument for causal exclusion is often accompanied by the following sort of diagram. M1 and M2 are singular tokens of particular mental events, and P1 and P2 are the corresponding physical event tokens on which the mental events depend.

$$\begin{tikzcd} M1 \arrow[r, dashed] \arrow[d, dash] & M2 \arrow[d, dash] \\ P1 \arrow[r, dashed] & P2 \end{tikzcd}$$

The vertical lines are of some kind of supervenience or ontological grounding relation between tokens of mental events and tokens of physical events. According to physicalism, all tokens of the mental must also be physical tokens. The horizontal line connecting P1 and P2 is causal: This represents the complete physical cause of P2. The question of mental causation is then the question of the causal efficacy of M1. There is an apparent causal relationship between M1 and M2. But, M1 supervenes on P1, and M2 supervenes on P2. P1 is a cause of P2, which means that there is already a complete cause for M2 in P1. Kim's claim is that M1 reduces to P1 as a cause: There is nothing about M2 that is leftover or uncaused, such that M1 can add by causally contributing to it. M1 does not cause P2, since the completeness of the physical means that P2 already has a sufficient physical cause, P1. What is left for M1 to do, asks Kim? M1 simply reduces to P1 in this picture, which means that the project of salvaging both physicalism and autonomous mental efficacy has failed.

The causal exclusion problem can be generalized to any set of token relata such that one kind of relata is identical with, supervenes on, is instantiated by, or is otherwise dependent on another kind of relata (see, for instance, Shapiro and Sober 2007). In other words, any so-called higher level phenomena will be subject to similar concerns about their genuine causal efficacy vis-à-vis their lower level counterparts. Evolutionary fitness, for instance, would, according to the generalized exclusion problem, have no causal efficacy of its own, being merely a stand-in for the causal efficacy of its lower level instantiations in individual instances of reproductive success. In the case of evolution, this may seem like the appropriate stance to take about the potential causal efficacy of higher level causes: We should refrain from reifying them all as having some additional mysterious kind of causal efficacy all their own. But in other cases, such as causal efficacy attributed to organisms as reproductive units, rather than, for instances, their genes, this is not as clearly the right result.

The generalized causal exclusion problem is the subject of criticism for the way in which it flattens causation, with the possibility of eliminating it altogether (see Block 2003). If the causal efficacy of higher level causes depends on or reduces to that of the lower level causes, then we can proceed to ask about the status of those lower level causes. Unless they are metaphysically bedrock, then they are in turn higher level than some other, yet-lower level causes. In this fashion, we proceed down to the microphysical, where we confront a new kind of dilemma. On one hand, we could commit to there being some lowest level such that only events at this level are genuinely causal. On the other hand, if there is no such lowest level, then causation "drains away" (Block 2003) into the bottomless abyss of the microphysical, and there is no genuine causal efficacy to be found anywhere.

The debate about causal drainage is somewhat off-topic for the issue of mental causation, but it helps keep a perspective on the urge to reduce higher level to lower level causes. While this may seem like a clearly justified maneuver in the context of a single pair of causal types (such as mental to physical, or psychological to neurophysiological), it requires making an assumption about what kinds of causes lack genuine efficacy, namely, those that are higher level with respect to other causes. This assumption may undercut the validity of the reduction by rendering those lower level causes just as inefficacious as the higher level ones. If neither the higher nor the lower have real causal efficacy, the impetus to reduce one to the other is greatly diminished.

There is some traction to be gained on the issue of causal exclusion and mental causation by considering the same structural problem for causal efficacy with different types of causes. Philosophers and scientists tend to have deeply entrenched views about what the "right" answer should be about mental causation, and this makes it easy to construct ad hoc "solutions" that have little merit other than yielding the correct conclusion, or to reject potentially viable accounts, because they get the wrong conclusion. By considering the same question applied to different relata, such as evolutionary fitness or thermodynamic temperature, we can assess different proposals reconciling or rejecting higher level causal efficacy on more neutral territory, and only then apply the solution to mental causation in particular.

The causal exclusion problem, applied to specifically mental causal relata, is the primary and dominant contemporary problem for mental causation. What is causing what, exactly, when we identify apparently mental causes, or offer causal explanations that rely on causal relationships involving mental relata?

Salvaging Mental Causation by Solving or Dissolving the Exclusion Problem

More solutions for solving the causal exclusion problem have been proposed in the last two decades than are possible to canvas in one place. The responses to this contemporary version of the problem of mental causation can be helpfully categorized in terms of the strategies they employ, however. Considering responses in

terms of these strategies is a useful way to map out the territory of ideas involved in the tangle of mental causation, mind-body relationship, reduction, explanation, and more.

Two popular general strategies are that of (i) changing the metaphysical type of the physical and mental tokens that are the relata in the above diagram, or (ii) changing the relationship posited between those tokens, such that the causal exclusion problem can no longer be formulated. Additional strategies involve various ways to bite the bullet, accepting that there is limited or no causal efficacy to the mental, and either (iii) rejecting the intuition that the mental should be treated as having causal efficacy, or (iv) offering a watered-down form of relevance for the mental that falls short of genuine causal efficacy but offers something more than pure epiphenomenalism. Two additional strategies dissolve the problem, rather than solving it directly: (v) denying the identity of mental and physical tokens, which means denying the relationship represented vertically in the above diagram, and (vi) challenging the implicit assumptions about causation on which the causal exclusion problem relies, with recourse to specific theories of causation. I will briefly discuss each of these strategies in order.

Strategy (i) starts by noting that the relata of mental and physical tokens used in formulating the causal exclusion problem could be of numerous different metaphysical types. A common approach is to follow Davidson and treat them as singular events, which is also convenient, since events are a very common relata type in theories of causation. Token mental events just are specific, very complicated, physical tokens, and the mental event tokens stand in all and only the causal relations in which the physical event tokens stand. Token-identity of events, however, leads straight to the causal exclusion problem whereby mental events do not have genuine causal efficacy. There are multiple potential candidates for the metaphysics of these tokens, but what unites these different accounts is that they each reconfigure the relata in the diagram above, as a way of attempting to block the reduction of the mental to the physical.

Instead of treating the mental and physical relata as events, one could construe them in terms of properties. Perhaps the mental is a second order property, i.e., a property had by some other, first order, property. If the mental is a second order property of first order physical properties, then it does not reduce away – it could be causally relevant or efficacious as a property of properties. There is thus only one token, having both physical, lower level, properties and mental, higher order, properties. Each instance of a mental property and a physical property are tokened by the same object. The mental property is multiply realizable, capable of being instantiated by a wide range of quite distinct physical tokens. As such, the mental properties do not reduce to the physical ones (see, on this, Bennett 2003; Levin 2010).

Strategy (ii) is similar to (i) in attempting to block causal exclusion by changing the characterization of the diagram above, but instead of focusing on the relata, the focus is on the relationships between them. The horizontal lines are causal; but what precisely is represented by the vertical lines? How one cashes out the relationship between the mental and the physical will affect the reducibility of mental causation

to physical causation. A common approach is that of some kind of supervenience of the mental on the physical (see especially Wilson 2005).

Supervenience can be a very broad asymmetric relationship, committing one to the claim that there can be no change in the mental without there also being a change in the physical, while there could potentially be a change in the physical without a change in the mental. For global supervenience, the mental parts of the world (be they events, properties, tropes, etc. – see strategy (i)) supervene on the entire physical world, such that any change, no matter how insignificant, might be sufficient for an entirely different set of mental relata to supervene. If we are considering the possible causal efficacy of Alice's intention to surprise Bob on the physical outcome of Bob's startled jump, then it is extremely counterintuitive to commit to the claim that the mental cause has changed because of a slight rearrangement of particles in a distant galaxy. Surely the mental relata supervene on a somewhat smaller batch of the physical world. Picking out what, exactly, goes into the patch of the physical world on which a particular candidate mental cause supervenes has proven to be a contentious issue.

Strategy (iii) involves denying the intuition that there is autonomous mental causation. Reductive physicalism of any stripe is an example of this, of which Kim himself is a well-known proponent. There are many versions of reductive physicalism, where the basic premise is that whatever the mental is, it reduces to or is nothing over and above the physical. Even stronger views, like eliminative physicalism (Churchland 1981), advocate the eventual eschewal of all mental terminology.

Strategy (iv) takes a somewhat different tack. Instead of simply denying the legitimacy of the intuition, some philosophers offer a replacement for genuine causal efficacy, something that explains why it seemed as if the mental were causally efficacious without having to grant full-blown mental causation. Jackson and Pettit (1990), for instance, distinguish between causal efficacy and causal relevance. Causal efficacy is what actually causes things to happen; causal relevance is what properties have such that they are cited in good causal explanations. But, while causally relevant properties may be explanatory, they lack causal efficacy. For example, according to this view, being fragile is causally relevant to the vase breaking, even though it was not the fragility that was causally efficacious in the actual breaking. This applies to the debate regarding mental causation by characterizing mental properties as causally relevant, even while it is always some other, quite specific, set of physical properties that are causally efficacious.

Strategy (v) is related to (i and ii) in that it focuses on the character of the relata and relationships in the diagram above, but instead of changing the kind of relationship between the mental and the physical, it argues that the vertical lines in the diagram do not exist. In other words, it denies the identity of token mental and physical relata, and by doing so, prevents the causal exclusion problem from being formulated. If the token of a mental event is not identical to or does not supervene solely on the token of a physical event, then the very way in which the problem is framed is specious (Andersen 2009).

There is a delicate balance to be struck in such strategies for responding to the causal exclusion problem for mental causation. The relationship between the mental and physical is itself as much a live issue for discussion, and as unresolved, as the existence or nature of mental causation itself. On the one hand, if we are willing to assume that the mental does have genuine causal efficacy, then we can use the causal exclusion problem as a rough guide toward how best to represent the relationship between the mental and physical. It would serve as a constraint on our representations that they yield the correct answer with respect to mental causation, and this would rule out some construals, such as token-identity of mental and physical events. It would not constrain enough to yield a single unequivocal solution, but it would be a substantive guide. On the other hand, though, we should be legitimately concerned about making this assumption about genuine mental causal efficacy, since we might take the causal exclusion problem to show exactly that this assumption is unwarranted. In that case, it is ad hoc to gerrymander our characterizations of the mental and the physical in order to reach the conclusion that the mental is genuinely efficacious. Rather than guiding us toward solving other problems like that of the mind-body relationship, it begs precisely the question at issue.

This leads to strategy (vi): Focus on causation as a way to situate mental causation in a broader perspective with better evidential footing. This tactic allows us to pose the question of whether or not, and in what way, the mental has genuine causal efficacy without having to make assumptions about how to best represent the relationship between the mental and physical relata in question. Accounts of causation provide independent evidential criteria for what it takes to count as a cause of something else; we can use these to investigate whether any mental "causes" actually meet the criteria for causation.

One way in which this has been done is to challenge the treatment of the mental and physical relata as sufficiently distinct that we can even sensibly ask the question about which one does the causing (Dardis 1993; Campbell 2010; Andersen 2009). Kim's original question, of whether it is M1 or P1 that causes M2, is misleading in that it asks us to implicitly treat these as distinct causal relata. They simply are not in competition, such that one – the physical – "wins out" over the other. Arguing from causal exclusion to the inefficacy of the mental is, on this strategy, a kind of category mistake.

Another way that strategy (vi) can be implemented to undermine the causal exclusion problem is to turn to contemporary accounts of causal explanation and see how mental causal relata fare on such accounts. Davidson's assumption about causation requiring general laws is widely rejected, and the causal exclusion problem cannot be formulated within many contemporary accounts of causation. In the last several decades, remarkable progress has been made in developing sophisticated philosophical and mathematical techniques for analyzing causal structure. We can translate the question of mental causation into specific accounts of causation to see if there is anything genuinely causal about the mental in each account. Once we do this, it turns out that on almost every contemporary account of causation, the mental has as much causal efficacy as any other cause outside of fundamental physics.

Consider an example of this. On the influential interventionist account of causation (Woodward 2003), the question of whether or not the mental has causal

efficacy is made more specific by considering a variety of causal variables that could represent different aspects of mental causation. One could ask about the role of conscious visual awareness in actions like reaching for and grasping objects. Each of these would be treated as a variable that takes on different values: perhaps Conscious visual experience (yes, no) and Reaching (hand preformed, hand not preformed). One would then consider the empirical research on this in order to determine if these two variables meet the criteria for having a causal relationship between them. Put very simply, this is a way of asking whether or not intervening on conscious visual awareness of an object changes the way in which we reach for those objects. Once we establish the answer (which, in this case, is yes – see Andersen 2009, Chap. 3), we have shown that there is at least one case of mental causation, namely, that which is represented by the variables in question.

Against this kind of mathematically sophisticated and scientifically grounded account, Kim's causal exclusion argument looks rather simplistic and naïve. One might want to protest that the variables are just a stand-in for the real causal story, which surely involves just the neurophysiological processes initiated by light impinging on the retina, and so forth. To defend this response in the face of a well-validated account of causation that has independent justification for its requirements on what counts as a cause, one should have more than just the intuition that the "real" causal story is elsewhere. This is a key advantage to treating the problem of mental causation as one of causation in general, rather than one of mental causes as sui generis: Given such well-established and explicitly justified methods for finding causal structure, it takes a great deal to show that the answer yielded by accounts of causation such as interventionism is incorrect. This puts the onus on reductive physicalists to show what is wrong in such analyses and to offer a more substantive defense of what counts as a "real" causal story, a very challenging task.

These six different argumentative strategies cover a vast number of different accounts, organized in terms of commonalities they share with regard to the aspect of the causal exclusion problem that they address. While there are multiple aspects to the issue of mental causation, there is no doubt that this problem, reconciling genuine mental efficacy with a physicalist view of the world and an ambiguous relationship between the mental and physical, is one of the main debates in contemporary philosophy of mind.

Conclusion and Future Directions

The causal exclusion problem is interesting for the way in which it captures so clearly the dilemma for genuine mental causation. The generalized exclusion problem raises issues for many different kinds of higher level causes, including the mental as well as a huge host of other potential causal relata from various sciences. However, the version of causal exclusion that applies to the issue of mental causation has, as we have seen, a unique twist that renders it particularly difficult to deal with. There is no clear answer to the question of how the levels in question are related to one another.

It is often tempting to treat mental causes as simply stand-ins for the "real" causal story, which must be purely physical and involve neurophysiological processes. After the last section, we should be cautious about this practice. Once we meet the evidential requirements for independently justified accounts of causation for the mental to have genuine causal efficacy, it is unnecessary to reject mental causation by claiming that "really" what we have just shown is somehow false or merely apparent. Furthermore, it is a substantive task to translate one potential relatum into another. Moving too quickly from the mental to the neurophysiological risks changing the subject, where the new physical process or event may be causally efficacious, but not efficacious *of the same effect* as was the original mental relatum.

This is where much of the future work regarding mental causation may be directed: toward a careful and detailed elaboration of the variety of ways in which mental causal relata can be translated into a format such that accounts of causation, coupled with results from cognitive neuroscience and psychology, can yield specific answers. The causal exclusion problem pitches mental causation as an all or nothing problem: Either it is causal or it is not. Moving away from this approach means moving toward a scientifically enriched process using well-confirmed tools from studies of causation. The goal then becomes to suss out the structure of mental causation and how it is situated in larger structures of the body and the environment.

Davidson was onto something when he said that mental causation, whatever it was, needed to be the same ordinary sort of causation that is studied in the sciences. It cannot be something different or new without thereby bringing in the sorts of mysterious powers that anyone with a commitment to physicalism of any stripe should want to avoid. He was wrong about what that kind of causation is – there is no need to assume unknowable universal laws governing physical redescriptions of mental causal relata. The way forward in this debate will be through foregrounding the issue of causation, making the question of mental causation more explicitly one of mental *causation*.

Cross-References

- ▶ Consciousness and Agency
- ▶ Determinism and Its Relevance to the Free-Will Question
- ▶ Explanation and Levels in Cognitive Neuroscience
- ▶ Free Will, Agency, and the Cultural, Reflexive Brain
- ▶ Realization, Reduction, and Emergence: How Things Like Minds Relate to Things Like Brains

References

Andersen, H. (2009). *The causal structure of conscious agency*. University of Pittsburgh. http://d-scholarship.pitt.edu/9254/. Pittsburgh, PA: University of Pittsburgh.

Bennett, K. (2003). Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs, 37*(3), 471–497.

Block, N. (2003). Do causal powers drain away? *Philosophy and Phenomenological Research, 67*(1), 133–150.
Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues, 20*(1), 64–79.
Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67–90.
Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind/brain*. Cambridge, MA: MIT Press.
Dardis, A. (1993). Sunburn: Independence conditions on causal relevance. *Philosophy and Phenomenological Research, 53*(3), 577–598.
Davidson, D. (1980/2001). *Essays on actions and events*. Oxford: Oxford University Press.
Davidson, D. (1995). Laws and cause. *Dialectica, 49*(2–4), 263–280.
Descartes, R. (1641/1996). *Meditations on the first philosophy* (trans: Cottingham, J.). Cambridge: Cambridge University Press.
Horgan, T. (1989). Mental quausation. *Philosophical perspectives*, *3*, 47–76.
Jackson, F., & Pettit, P. (1990). Program explanation: A general perspective. *Analysis, 50*(2), 107–117.
Kim, J. (1989). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association, 63*(3), 31–47.
Kim, J. (2000). *Mind in a physical world Cambridge*. Cambridge, MA: MIT Press.
Levin, J. (2010). Functionalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 ed.). http://plato.stanford.edu/archives/sum2010/entries/functionalism/
Robb, D. & Heil, J. (2013). Mental causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2013 ed.). http://plato.stanford.edu/archives/spr2013/entries/mental-causation/
Shapiro, L., & Sober, E. (2007). Epiphenomenalism: The dos and don'ts. In G. Wolters & P. Machamer (Eds.), *Thinking about causes: From Greek philosophy to modern physics* (pp. 235–264). Pittsburgh: University of Pittsburgh Press.
Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review 68*(2), 141–156.
Wilson, J. (2005). Supervenience-based formulations of physicalism. *Noûs, 39*(3), 426–459.
Woodward, J. (2003) Making Things Happen. New York, NY: Oxford University Press.

Neural Representation and Computation

6

Corey J. Maley and Gualtiero Piccinini

Contents

Abstract

Nervous systems perform amazing control functions, which include driving complex locomotive systems in real time. How do they do it? The best explanation neuroscientists have found is that nervous systems collect information from the organism and the environment, use that information to construct representations, and perform computations on such representations. The output

C.J. Maley (✉)
Department of Philosophy, Princeton University, Princeton, NJ, USA
e-mail: cmaley@princeton.edu

G. Piccinini
Department of Philosophy, University of Missouri – St. Louis, St. Louis, MO, USA
e-mail: piccininig@umsl.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_8,

of neural computations drives the organism. This article discusses what it means for nervous systems to carry information, to represent, and to perform computations.

Neural Information

This section discusses information, a ubiquitous notion which turns out to have multiple senses. Particularly when dealing with neural systems, it is important to distinguish what kind of information is being discussed. The most relevant notions are Shannon information – a precise, mathematical notion often used in characterizing neural systems – and natural semantic information, an informal but useful notion of information that lets us begin to assign meanings to neural events.

Neural Information I: Shannon Information

In his mathematical theory of information, Shannon (1948) defined two quantitative measures of information, *entropy* and *mutual information*. Shannon entropy is a quantitative, objective measurement of how informative a state change is in a system. Very generally, if a system goes from one state to another, the informational "value" of the second state is greater if the chances of moving to that state are very low; if it is very likely that the system move to that second state, it is less informative.

More formally, entropy can be defined as follows. Let X be a discrete random variable such that X takes values in $\{x_1, x_2, \ldots x_n\}$ with respective probabilities $p(x_1), p(x_2), \ldots p(x_n)$. We assume that these probabilities are all greater than zero, and that the sum of the probabilities is equal to one. The entropy of the random variable is defined as follows:

$$H(X) = -\sum_{j=1}^{n} p(x_j) \log_2 p(x_j).$$

The entropy tells us how much information is transmitted by a random variable. So, for example, if X represents flips of a fair coin, then it takes values of heads or tails with probability 0.5. The entropy in that case is 1, which means that exactly one "bit" of information is gained by each transmission/reception of X. If heads becomes more probable than tails (as in the case of a loaded coin), however, then the entropy decreases, which coincides with the intuitive idea that less information is gained if we know that heads is more likely. If we take the extreme case in which heads is certain, then the entropy is zero: if we already know that the coin must be heads, learning that it is heads on a particular flip makes no difference. Another way to view entropy is as a measure of uncertainty about the state of a system: a decrease in a system's entropy means we are more certain about what state (or states) the system will be in. In our coin example, as entropy approaches zero, the uncertainty of its landing heads decreases.

The second quantitative measure Shannon introduced is called mutual information. Informally, mutual information is a measure of how much we can determine about one variable given another variable. If two variables are independent, then nothing can be known about one by observing the other, and their mutual information is zero. But non-zero mutual information about two variables implies that the state of one variable makes it more likely that the second variable is in certain states than others.

Formally, mutual information is given by:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)$$

As before, X is a discrete random variable, and so is Y. The term $p(x, y)$ is the joint probability of X and Y. Note that mutual information is *mutual*: non-zero mutual information between X and Y is precisely the same between Y and X.

One of the most immediate and important uses of information theory is known as coding theory. Here, information theory is used to analyze and design communication systems. In real-world systems, noise of some sort is introduced into the signal going from the sender to the receiver, and the central task of coding theory is to design ways of encoding information such that the intended message is retrievable in the presence of such noise. More important for the present discussion is the application of information theory to neural signaling, in which neural systems are analyzed using the tools of coding theory.

Using information theory in this way, neuroscientists can study neural signaling systems. It is uncontroversial that neural firings – action potentials, or excitatory postsynaptic potentials (EPSPs) – are the primary means by which neural communication occurs, and which thus undergirds the sensory, cognitive, and motor processes for which the brain is responsible. One example comes from Borst and Theunissen (1999): here, the authors use coding theory to analyze the information content carried by spikes from individual neurons according to different models of neural spiking. Furthermore, they demonstrate that information theory can be used to show that so-called temporal codes, in which the relative time between individual spikes is thought to carry information, do not seem to be a good model for how information is transmitted in most neural systems.

In summary, Shannon entropy and mutual information play an important role in analyzing neural systems – much can be learned about the ways that neurons generate action potentials and respond to these signals by applying Shannon's quantitative measures of information – but only in a limited way. For neither entropy nor mutual information tell us what any particular signal means – what any particular signal tells us about any particular state of affairs. Whenever there is mutual information between a source and a receiver, by definition the receiver obtains information *about* the sources – that is, the receiver learns something about what happens at the source. This can be the basis for a useful notion of semantic information.

Neural Information II: Natural Semantic Information

When we want to know what a particular alarm means, or what a word in a foreign language means, we are certainly asking about the information conveyed in each case, but not in the sense of Shannon information. We call this kind of information "semantic information," adopt Grice's influential distinction between "natural" and "non-natural" semantic meaning (Grice 1957), and apply it to natural information.

Natural semantic information is the information conveyed about a state of affairs by way of some natural process. A paradigmatic example is the information gained when one sees a column of smoke rising in the forest: the smoke carries the information that there is a fire, and so seeing smoke allows us to infer that there is a fire near the base of the smoke column. Smoke carries natural information about fire because it reliably correlates with fire. Thus, we will say that a signal carries natural semantic information about what it reliably correlates with.

There are many examples of natural semantic information in neural systems. A few examples include place cells, grid cells, and feature detectors. Place cells are hippocampal neurons that fire preferentially when an organism is in a particular location in its local environment: a particular place cell fires most vigorously when the animal is in a particular place. One place cell might fire when the organism is in a particular corner of its enclosure; another might fire when it's in a different corner. Grid cells function similarly, but instead of firing preferentially when the animal is in a particular *place*, grid cells fire at regular intervals within the local environment. If we were to map the grid cells that fire as an animal moves within its environment, a regular, triangular "grid" would become apparent, suggesting that the grid cell system tracks the animal's location relative to an imaginary grid laid out over its environment. The place cells, however, might only be sensitive to a small number of locations, such as particular "landmarks" within the environment, but not others (Moser et al. 2008).

The class of neurons known as feature detectors is another good example of a system that carries natural semantic information. Beginning with the seminal work of Hubel and Wiesel, neuroscientists have demonstrated that certain neurons in the visual system fire preferentially when the organism is viewing particular visual patterns, such as lines, angles, shapes, or images moving in a particular direction (Hubel and Wiesel 1962). So, for example, one neuron might fire most vigorously when the animal is viewing a horizontal line, while another fires most vigorously when the animal is viewing a vertical line. By combining this information, neurons "upstream" from these basic feature detectors form more complex feature detectors, able to detect, for example, the edges of objects and environmental scenes (Marr and Hildreth 1980).

These are all examples in which neuroscientists have experimentally demonstrated reliable correlations between environmental features and the firing (or firing patterns) of neurons. These carriers of natural semantic information are obviously quite important to organisms: grid and place cells are part of the explanation for how animals navigate, and feature detectors are part of the explanation for how animals recognize aspects of their environment. Note again that this sense of

information is stronger than that of Shannon information: while neural systems involved in the processing of natural semantic information can be analyzed in terms of Shannon information, a system that can be analyzed in Shannon information-theoretic terms need not carry natural semantic information at all. We can measure the entropy of a neural spiking event without knowing (or caring) about what the spikes are *about*. For example, Strong et al. (1998) demonstrate how to calculate the entropy of signals from a neuron known as H1, based on the time between spikes. They mention that although neuroscientists understand what H1 represents, that has nothing to do with the analysis they are proposing, which is of H1's informational entropy.

This section outlined two different senses of information. Beyond mere conceptual clarity, it is useful to make clear which one of these two senses is being deployed in neuroscientific and philosophical discussions of the mind/brain. The claim that a particular neural system is processing information, or is being understood as an information processing system, may mean nothing more than that the system is being analyzed in terms of its Shannon-information properties; that is no guarantee that a system is capable of carrying semantic information. Furthermore, characterizing a system as capable of processing information is no guarantee that the system manipulates – or should be understood as manipulating – representations.

Neural Representation

Having distinguished different senses of information gives us the tools we need to make clear when precisely an instance of neural information-carrying is an instance of neural representation. The standard account of what a representation is will not do for neuroscience, but there are kinds of neural information-carrying that are instances of neural representation in a sense that serves neuroscience well.

The Standard Conception of Representation

The received view of representation may be found in C. S. Peirce's seminal work on what he called signifiers (Peirce 1992). Using contemporary terminology, representations consist of three essential aspects. First, there is the thing to be represented, often called simply the *object*. Next, there is the thing that represents that object: the *representation* (Peirce calls this the *signifier*). Finally, there is the *subject* or *agent* who interprets the representation as being about, in some way or another, the object (Peirce calls this the *interpretant*). A simple example might be a drawing of a cat. The object is the cat itself, the representation is the drawing, and the agent is whoever sees the drawing as being about the cat.

There are several difficulties with the standard conception of representation. One difficulty is how to understand the "being about" relation mentioned above. Peirce noted that there are three ways in which representations are about their objects.

Representations can be correlated with their objects (compare with the above discussion of natural semantic information): smoke can represent fire because smoke is caused by fire. Representations can also resemble their objects: a picture of a fire is about a fire because it resembles a fire, at least in its visual aspects. Last, representations can be about objects by convention: the word "fire" is about fire because this is the word that English language speakers agree to use for this purpose.

There remains a problem with the role of the agent (interpretant). Peirce's three-place account works quite well for everyday examples of representation. But it won't work for an account of representation appropriate to neuroscience. To see why, consider a simple example, such as the claim that certain patterns of activation in a particular neural system represent faces. The object and representation are clear: they are a particular face, and a particular pattern of activation, respectively. But what is the interpretant? It can't be the whole organism, because organisms do not have that kind of access to the activity of their neural systems. Positing that some subsystem of the organism is the interpretant amounts to equating that subsystem with a homunculus: another agent with the representational capacities to recognize that the representation is about the object. The problem with homuncular explanations is clear: what we are trying to explain by positing representations is some capacity of the organism, and positing homunculi with the very capacities we are trying to explain is no explanation at all. It can't be homunculi all the way down!

Thus, what's needed is a conception of representation that both avoids the problem of homuncular regress and warrants being taken seriously as a conception of *representation*, rather than something else.

A Functional Conception of Representation

The problem with standard accounts is that an interpretant is necessary to make sense of what a representation is about. An alternative account eliminates the role of the interpretant and posits that a representational system's *function* can undergird the fact that representations within that system are about particular objects (e.g., Dretske 1988). This allows us to understand representations without positing homunculi.

So how does a functional account accomplish this? Consider the example of a gas gauge that shows the amount of gas in a car by the weight of the tank. The gauge indicates a number of things: the weight of the gas, the force due to gravity being exerted on the bolts holding the gas tank to the car, and so on. While the gauge carries natural semantic information about all of these things, it only *represents* one thing: the amount of gas in the tank. This is because the function of the gauge is to carry natural semantic information about the amount of gas, but not the gravitational force on certain bolts. This function is important for the control of certain behaviors, such as illuminating a low-fuel light, and alerting the driver when it is time to get more gas.

This idea extends usefully to natural systems: the pattern of activation within a neural system, for example, might *indicate* many things, such as the amount of neurotransmitter or mean electrical activity in a given region, but it only *represents*, say, the angle of the organism's limb. This pattern of activity represents limb angle because carrying natural semantic information about limb angle is the *function* of the neural system, and not because there is anyone – a homunculus or an outside observer – interpreting those neural firings as being about limb angle.

This account relies heavily on the notion of function. Scientists – particularly those working in the special sciences – appeal to this notion of function with some frequency, and a number of philosophers have gone some way toward providing accounts of biological functions that are both philosophically satisfactory and do justice to the practices of working scientists (some examples include (Allen and Bekoff 1995; Bigelow and Pargetter 1987; Godfrey-Smith 1994; Millikan 1989)).

Some Neural States are Representations

Everything that was said above can be put together to make clear when a neural state, or a part of a neural system, is genuinely a representation, or processing representations. To reiterate, it is not enough that a neural system carries information; the kind of information being carried matters. Trafficking in Shannon-information is not enough; the system must carry natural semantic information. But carrying natural semantic information is still not enough to generate a representation: tree rings may carry natural semantic information about the number of times the earth has revolved around the sun since the birth of the tree, but those rings do not represent anything at all: they do not have the function of representing. What is required is that natural semantic information is carried or processed in a way such that the system containing that information has the function of carrying that information.

A specific example from neuroscience may be useful at this point. Romo and colleagues describe several cases involving monkeys in which researchers have demonstrated the correlation between particular neural systems and the stimulation of the monkeys' skin (Romo et al. 2002). The particular stimulation involved is vibrotactile: the stimulus vibrates on the surface of the monkeys' fingers at variable frequencies and intensities, which the experimenters control. A number of experiments have demonstrated that there are direct correlations between the patterns of activity in particular neurons in the primary somatosensory cortex (called S1) and the particular kind of stimulation applied to monkeys' fingers; thus, these neurons carry natural semantic information about an external stimulus. More interestingly, the monkeys can be trained to perform tasks that require discriminating particular patterns of stimulation from other such patterns; this strongly suggests that the function of these neurons in S1 is to detect this kind of stimulation. If so, these neurons represent the pattern of stimulation on the monkeys' fingers, and the information carried by this representation is used in other parts of the monkeys' cortex in discrimination tasks that are sensitive to this information.

The above sections have shown that information processing and representation are two separate things: although representation involves carrying information, information-carrying may not involve any representation at all. The final section addresses how these notions relate to computation in neural systems.

Neural Computation

Computation is often considered the same thing as information processing or representation manipulation (e.g., Fodor 1981). But while computation is often used to process information and manipulate representations, computation can also occur without information or representation (Piccinini 2008). Therefore, we offer an account of computation that is applicable to neuroscience but does presuppose representation or information processing. Such an account paves the way for assessing different versions of a doctrine known as computationalism, according to which the nervous system is a kind of computing system.

Computation: A Mechanistic Account

According to the mechanistic account of computation (Piccinini and Scarantino 2011), a computation in the most general sense is the processing of vehicles (state variables) according to rules that are sensitive to certain vehicle properties, and specifically, to differences between different portions of the vehicles. This processing is performed by a functional mechanism, which is a mechanism whose components are functionally organized to perform the computation (i.e., the organized components have the function of performing a computation). Thus, if the mechanism malfunctions, a miscomputation occurs.

When we define concrete computations and the vehicles that they manipulate, we need not consider all of their specific physical properties. Rather, we only need to consider the properties that are relevant to the computation as specified by the rules defining the computation. A physical system can be described at different levels of abstraction; since concrete computations and their vehicles are described at a level of abstraction defined independently of the physical media that implement them, we call them *medium-independent*.

To put the point another way, a vehicle is medium-independent when the rules (i.e., the input–output maps) that define a computation are sensitive only to differences between portions of the vehicles along specific dimensions of variation, and insensitive to any other concrete physical properties of the vehicles. Thus, a given computation can be implemented in different physical media (e.g., mechanical, electro-mechanical, electronic, magnetic, etc.), provided that the media possess a sufficient number of dimensions of variation (or degrees of freedom) that can be appropriately manipulated, and that the components of the implementing mechanism are functionally organized in the appropriate way.

Computationalism

Computationalism is the view that cognitive capacities have a computational explanation or, somewhat more strongly, that cognition just is a kind of computation. In what follows, these two formulations are used interchangeably. Many neuroscientists endorse some version of computationalism: when neuroscientists propose an explanation of a cognitive capacity, the explanation typically involves computations underlying the cognitive capacity. But neuroscientists and cognitive scientists differ on how they apply the notion of computation to the study of cognition. Their three main research traditions are classical computationalism, connectionism, and computational neuroscience.

The best computational theories appeal to a well-defined kind of computation. Historically, the most influential version of computationalism appeals to digital computation – the kind of computation performed by digital computers. Other versions appeal to analog computation or computation in a generic sense, which encompasses both digital and analog computation as species.

Three Research Traditions: Classical Computationalism, Connectionism, and Computational Neuroscience

The view that thinking has something to do with computation may be found in the works of some early modern materialists, such as Thomas Hobbes (Boden 2006). But computationalism properly began in earnest after a number of logicians (most notably Alonzo Church, Kurt Gödel, Stephen Kleene, Emil Post, and especially Alan Turing) laid the foundations for the mathematical theory of computation.

Turing (1936–7) analyzed computation in terms of what are now called Turing Machines (TMs) – a kind of simple, idealized processing unit operating on an unbounded tape. The tape is divided into squares, upon which the processor can write symbols or read existing symbols. The processor moves along the tape reading and writing on one square at a time. The processing unit's behavior – what symbol to write, where to move – is governed by rules; these rules specify what to do depending on what is on the tape as well as which of finitely many states the processor is in.

Turing argued convincingly that any function on the natural numbers that can be computed by following an algorithm (i.e., an unambiguous list of instructions operating on discrete symbols) can be computed by a TM. Church offered a similar proposal in terms of general recursive functions, and it turns out that a function is general recursive if and only if it can be computed by a TM. Given this extensional equivalence between TMs and general recursive functions, the thesis that any algorithmically computable function is computable by some TM (or equivalently, is general recursive) is now known as the Church-Turing thesis.

Turing made two other relevant contributions. First, he showed how to construct what are now called Universal Turing Machines (UTMs). These are TMs that can mimic any other TM by encoding the rules that govern that machine as instructions,

storing the instructions on a portion of their tape, and then using the encoded instructions to determine their behavior on the input data. To put this in more modern terms, UTMs can take in a description of any TM and use that description as a program. Notice that ordinary digital computers, although they have more complex components than UTMs, are universal in the same sense (up to their memory limitations). That is, digital computers can compute any function computable by a TM until they run out of memory.

Second, Turing showed that the vast majority of functions whose domain is denumerable (e.g., functions of strings of symbols or of natural numbers) are actually *not* computable by TMs. These ideas can be put together as follows: assuming the Church-Turing thesis, a universal digital computer can compute any function computable by algorithm, although the set of these Turing-computable functions is a tiny subset of all the functions whose domain is denumerable.

Modern computationalism began when Warren McCulloch and Walter Pitts connected three things: Turing's work on computation, the explanation of cognitive capacities, and the mathematical study of artificial neural networks. Artificial neural networks are sets of connected signal-processing elements ("neurons" or "nodes"). Typically, they have nodes that receive inputs from the environment (input elements), nodes that yield outputs to the environment (output nodes), and nodes that communicate only with other nodes in the system (hidden nodes). Each node receives input signals and delivers output signals; the output is a function of the received input and the node's current state. As a result of their nodes' activities and organization, neural networks turn the input received by their input nodes into the output produced by their output nodes. A neural network may be either a concrete physical system or an abstract mathematical system. An abstract neural network may be used to model another system (such as a network of actual neurons) to some degree of approximation.

The mathematical study of real neural networks using biophysical techniques began around the 1930s. McCulloch and Pitts were the first to suggest that neural networks have something to do with computation. They defined networks that operate on sequences of discrete inputs in discrete time. They then argued that their networks are a useful idealization of real neural networks, and concluded that the activity of their networks could explain cognitive phenomena. McCulloch and Pitts also pointed out that their networks can perform computations like those of TMs. More precisely, McCulloch-Pitts networks are computationally equivalent to TMs without tape (or finite state automata). Modern digital computers are a kind of McCulloch-Pitts neural network. Digital computers are sets of logic gates – digital signal-processing elements equivalent to McCulloch-Pitts neurons – connected to form a specific architecture.

McCulloch and Pitts's account of cognition contains three important aspects: an analogy between neural processes and digital computations, the use of mathematically defined neural networks as models, and an appeal to neurophysiological evidence to support their neural network models. After McCulloch and Pitts, many others linked computation and cognition, though they often abandoned one or more aspects of McCulloch and Pitts's theory. Computationalism evolved into three main traditions, each emphasizing a different aspect of McCulloch and Pitts's account.

One tradition, sometimes called classical computationalism, emphasizes the analogy between cognitive systems and digital computers while downplaying the relevance of neuroscience to the theory of cognition. When researchers in this tradition offer computational models of a cognitive capacity, the models take the form of computer programs for producing the capacity in question. One strength of the classicist tradition lies in programming computers to exhibit higher cognitive capacities such as problem solving, language processing, and language-based inference.

A second tradition, most closely associated with the term "connectionism" (although this label can be misleading; see below), downplays the analogy between cognitive systems and digital computers in favor of computational explanations of cognition that are "neurally inspired". When researchers in this tradition offer computational models of a cognitive capacity, the models take the form of artificial neural networks for producing the capacity in question. Such models are primarily constrained by psychological data, as opposed to neurophysiological and neuroanatomical data. One strength of the connectionist tradition lies in designing artificial neural networks that exhibit cognitive capacities such as perception, motor control, learning, and implicit memory.

A third tradition is most closely associated with the term computational neuroscience, which is one aspect of theoretical neuroscience. Computational neuroscience downplays the analogy between cognitive systems and digital computers even more than the connectionist tradition. The reasons for this are less straightforward than it seems, but it does appear that the functional relevance of neural signals depends on non-discrete aspects of the signals such as firing rates and spike timing. More specifically, current neuroscientific evidence suggests that typical neural signals, such as spike trains and spike timings, are graded like continuous signals but are constituted by discrete functional elements (spikes); thus typical neural signals are neither continuous signals nor strings of digits. Therefore, a strong case can be made that typical neural signals are neither continuous signals nor strings of digits, and that in the general case neural computation is neither digital nor analog but sui generis (Piccinini and Bahar 2013).

Neurocomputational models aim to describe actual neural systems such as (parts of) the hippocampus, cerebellum, or cortex, and are constrained by neurophysiological and neuroanatomical data in addition to psychological data. It turns out that McCulloch-Pitts networks and many of their "connectionist" descendents are relatively unfaithful to the details of neural activity, whereas other types of neural networks are more biologically realistic. Computational neuroscience offers models of how real neural systems may exhibit cognitive capacities, especially perception, motor control, learning, and implicit memory.

Although the three traditions just outlined are in competition with one another to some extent, there is also some fuzziness at their borders. Some cognitive scientists propose hybrid theories, which combine explanatory resources drawn from both the classicist and the connectionist traditions. In addition, biological realism comes in degrees, so there is no sharp divide between connectionist and neurocomputational models.

Understanding the Three Traditions

The debate between classicists and connectionists has been somewhat confusing. Different authors employ different notions of computation, which vary in both their degree of precision and their inclusiveness. Specifically, some authors use the term 'computation' only for classical computation – i.e., algorithmic digital computation over language-like structures – and conclude that connectionism falls outside computationalism. By contrast, other authors use a broader notion of computation, thus including connectionism within computationalism. But even after we factor out differences in notions of computation, further confusions can be easily found.

Classical computationalism and connectionism are often described as being at odds with one another for two reasons. First, classical computationalism is committed to the idea that the vehicles of digital computation are language-like structures. Second, classical computationalism is taken to be autonomous from neuroscience. Both of these theses are flatly denied by many prominent connectionists. But some connectionists also model and explain cognition using neural networks that perform computations defined over digital structures, so perhaps they should be counted among the digital computationalists.

Furthermore, both classicists and connectionists tend to ignore computational neuroscientists, who in turn tend to ignore both classical computationalism and connectionism. Computational neuroscientists often operate with their own mathematical tools without committing themselves to a particular notion of computation. To make matters worse, some connectionists and computational neuroscientists reject digital computationalism – they maintain that their neural networks do not perform digital computations.

In addition, the very origin of digital computationalism calls into question the commitment to autonomy from neuroscience. McCulloch and Pitts initially introduced digital computationalism as a theory of the brain, and some form of computationalism or other is now a working assumption of many neuroscientists. For example, Koch (1999) begins with these two sentences: "The brain computes! This is accepted as a truism by the majority of neuroscientists engaged in discovering the principles employed in the design and operation of nervous systems."

A further wrinkle derives from the ambiguity of the term "connectionism." In its original sense, connectionism says that behavior is explained by the changing "connections" between stimuli and responses, which are biologically mediated by changing the strength of the electrochemical connections between neurons. This original connectionism is a descendant of associationist behaviorism, so it may be called associationist connectionism. According to associationist behaviorism, behavior is explained by the association between stimuli and responses, that is, organisms learn through reinforcement to respond to stimuli in certain ways. Associationist connectionism adds to associationist behaviorism a biological mechanism to explain these associations: modifying the strength of the connections between neurons.

But contemporary connectionism is a more general thesis than associationist connectionism. In its most general form, contemporary connectionism, like

computational neuroscience, simply says that cognition is explained (at some level) by the activity of neural networks. This is a truism, or at least it should be. The brain is the organ of cognition, the cells that perform cognitive functions are (mostly) neurons, and neurons perform their cognitive labor via their activity in networks. Neural activity is computation at least in a generic sense. Because digital computers are a special kind of neural network, even classicists, whose theory is most closely inspired by digital computers, are committed to connectionism in its general sense.

The relationship between connectionist and neurocomputational approaches on one hand and associationism on the other turns on a distinction between strong and weak associationism. Strong associationism maintains that association is the only legitimate explanatory construct in a theory of cognition. Weak associationism maintains that association is a legitimate explanatory construct along with others such as the innate structure of neural systems.

To be sure, some connectionists profess strong associationism. But that is beside the point, because connectionism *per se* is consistent with weak associationism or even the complete rejection of associationism. Some connectionist models do not rely on association at all – a prominent example being the work of McCulloch and Pitts. Weak associationism is consistent with many theories of cognition, including classicism. A vivid illustration is Alan Turing's early proposal to train associative neural networks to acquire the architectural structure of a universal computing machine. In Turing's proposal, association may explain how a network acquires the capacity for universal computation (or an approximation thereof), while the capacity for universal computation, in turn, may explain any number of other cognitive phenomena.

Although many of today's connectionists and computational neuroscientists emphasize the explanatory role of association, many of them also combine association with other explanatory constructs, as per weak associationism. What remains to be determined is which neural networks, organized in what way, actually explain cognition and which role association and other explanatory constructs should play in a theory of cognition.

Yet another source of confusion is that classical computationalism, connectionism, and computational neuroscience tend to offer explanations at different mechanistic levels. Specifically, classicists tend to offer explanations in terms of rules and representations, without detailing the neural mechanisms by which the representations are implemented and processed; connectionists tend to offer explanations in terms of highly abstract neural networks, which do not necessarily represent networks of actual neurons (in fact, a processing element in a connectionist network may represent an entire brain area rather than an actual neuron); finally, computational neuroscientists tend to offer explanations in terms of mathematical models that represent concrete neural networks based on neurophysiological evidence. Explanations at different mechanistic levels are not necessarily in conflict with each other, but they do need to be integrated to describe a multi-level mechanism. Integrating explanations at different levels into a unified multi-level mechanistic picture may require revisions in the original explanations themselves.

Different parties in the dispute between classical computationalism, connectionism, and computational neuroscience may offer different accounts of how the different levels relate to one another. One traditional view is that computational explanations do not describe mechanisms. Instead, computational and mechanistic explanations are independent. This suggests a division of labor: computations are the domain of psychologists, while the implementing neural mechanisms are the business of neuroscientists. According to this picture, the role of connectionists and computational neuroscientists is to discover how neural mechanisms implement the computations postulated by (classicist) psychologists.

This traditional view has been criticized as unfaithful to scientific practices. It has been pointed out that, first, that neuroscientists also offer computational explanations (not just psychologists); second, far from being independent, different levels of explanation constrain one another; and finally, both computational explanations and mechanistic explanations can be given at different levels.

One alternative to the traditional view is that connectionist or neurocomputational explanations simply replace classicist ones. Perhaps some connectionist computations approximate classical ones. In any case, some authors maintain that classicist constructs, such as program execution, play no causal role in cognition and will be eliminated from cognitive science.

A more neutral account of the relation between explanations at different levels is provided by the mechanistic account of computation. According to the mechanistic account, computational explanation is just one type of mechanistic explanation. Mechanistic explanations provide components with such properties and organization that they produce the phenomenon. Computational explanation, then, is explanation in terms of computing mechanisms and components – mechanisms and components that perform computations. Mechanistic explanations come with many levels of mechanisms, where each level is constituted by its components and the way they are organized. If a mechanistic level produces its behavior by the action of computing components, it counts as a computational level. Thus, a mechanism may contain zero, one, or many computational levels, depending on what components it has and what they do. Which types of computation are performed at each level is an open empirical question to be answered by studying cognition and the nervous system at all levels of organization.

Conclusion

Computationalism is here to stay. There is every reason to suppose that cognitive capacities have computational explanations, at least in a generic sense. Moreover, everyone agrees with (or should agree with) connectionists or computational neuroscientists, at least in the general sense of embracing neural computation over neural representations. Nonetheless, much work remains to be done.

A complete computational study of cognition will require that we integrate different mechanistic levels into a unified, multi-level explanation of cognition. We also need to characterize the specific computations on which cognition

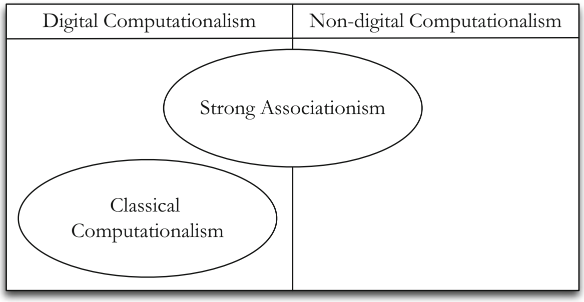

Fig. 6.1 Some prominent forms of computationalism and their relations

depends: whether – and to what extent – the satisfactory explanation of cognition requires classical computational mechanisms as opposed to non-classical digital computation, and whether we need to invoke processes that involve non-digital computation (Fig. 6.1). It may turn out that one computational theory is right about all of cognition, or it may be that different cognitive capacities are explained by different kinds of computation. To address these questions in the long run, the only effective way is to study nervous systems at all its levels of organization and find out how they exhibit cognitive capacities.

Cross-References

► Experimentation in Cognitive Neuroscience and Cognitive Neurobiology
► Explanation and Levels in Cognitive Neuroscience

References

Allen, C., & Bekoff, M. (1995). Biological function, adaptation, and natural design. *Philosophy of Science, 62*(4), 609–622.
Bigelow, J., & Pargetter, R. (1987). Functions. *The Journal of Philosophy, 84*(4), 181–196.
Boden, M. (2006). *The mind as machine*. Oxford: Oxford University Press.
Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature Neuroscience, 2*(11), 947–957. doi:10.1038/14731.
Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
Fodor (1981). *Scientific American, 244*, 114–25.
Godfrey-Smith, P. (1994). A modern history theory of functions. *Noûs, 28*(3), 344–362.
Grice, H. P. (1957). Meaning. *Philosophical Review, 66*(3), 377–388.
Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*, 106–154.
Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. Oxford: Oxford University Press.
Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society B, 207*, 187–217.
Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science, 56*(2), 288–302.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience, 31*(1), 69–89. doi:10.1146/annurev.neuro.31.061307.090723.

Peirce, C. S. (1992). In N. Houser & C. Kloesel (Eds.), *The essential peirce*. Bloomington: Indiana University Press.

Piccinini, G. (2008). Computation without representation. *Philosophical Studies, 137*(2), 205–241.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science, 34*, 453–488.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics, 37*(1), 1–38. doi:10.1007/s10867-010-9195-3.

Romo, R., Hernandez, A., Zainos, A., Brody, C., & Salinas, E. (2002). Exploring the cortical evidence of a sensory-discrimination process. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 357*(1424), 1039–1051. doi:10.1098/rstb.2002.1100.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.

Strong, S. P., Koberle, R., van Steveninck, R. R. D. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters, 80*(1), 197.

Turing, A. M. (1936–7 [1965]). *On computable numbers, with an application to the Entscheidungsproblem* (Repreinted in *The Undecidable*, pp. 116–154, by M. Davis, Ed., Ewlett, Raven).

Section II

Moral Cognition

Moral Cognition: Introduction

7

Stephan Schleim

Contents

Abstract

Research on moral cognition is a growing and heavily multidisciplinary field. This section contains chapters addressing foundational psychological, neuroscientific, and philosophical issues of research on moral decision-making. Furthermore, beyond summarizing the state of the art of their respective fields, the authors formulate their own proposals to answer open questions such as those on the relation between emotion and cognition in moral psychology, the idea that there is a "moral module" in the human brain, the relevance of this research for ethics and meta-ethics, the various psychological and philosophical meanings of "intuition" and how intuitions can have a justificatory role, or the connection between the psychological, neuroscientific, and philosophical levels in popular experiments on moral cognition. Research on moral decision-making is challenging, for empiricists as well as theoreticians, and is related to several applied questions of neuroethics which are briefly addressed at the end of this introduction.

S. Schleim
Faculty of Behavioral and Social Sciences, Theory and History of Psychology, Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

Neurophilosophy, Munich Center for Neurosciences, Ludwig-Maximilians-University Munich, Munich, Germany
e-mail: s.schleim@rug.nl

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_94,

Introduction: Moral Cognition

The early twenty-first century has seen an increasing academic interest in moral cognition that has been truly multidisciplinary, connecting branches of philosophy with empirical disciplines, reconsidering traditional questions of ethics or meta-ethics – for those unfamiliar with the term: "the attempt to understand the metaphysical, epistemological, semantic, and psychological, presuppositions and commitments of moral thought, talk, and practice" (Sayre-McCord 2012) – as well as age-old debates in the light of new scientific evidence (see Fig. 7.1). To name just two broad and important examples which are stimulated by new psychology and neuroscience findings and which are also intensively discussed in the contributions to this section: First, the question of the nature and role of moral reasoning, emotion, or intuition and, second, the question of whether there are natural moral facts and properties, perhaps even "hard-wired" into our brain, as once suggested by Michael Gazzaniga (2005).

Philosophers have debated whether moral judgments *should* be based in the passions or in reason, with the Scottish philosopher David Hume being a famous proponent of the passions (Hume 1777/1975) and the German philosopher Immanuel Kant an eminent advocate of reason (Kant 1785/2011; for a compilation of historical readings, see Nadelhoffer et al. 2010). Sometimes inspired by that traditional historical debate but with a different question in mind, psychologists tried to understand *how* moral judgments *are actually made* by lay people or certain kinds of experts, sometimes with an emphasis on the development of their moral faculties as they mature from childhood to adulthood or as a function of their education (Gilligan 1982; Kohlberg 1984; Lind et al. 2010; Piaget 1932), sometimes focusing on situational characteristics of a particular environment influencing moral judgments (Haidt 2001; Pizarro 2000).

With the expansion of cognitive neuroscience and the neuroscience turn in psychology taking place (Littlefield and Johnson 2012; Pickersgill and van Keulen 2012), it was probably just a question of time before moral judgments were investigated in a functional magnetic resonance imaging brain scanner (Greene et al. 2001; Moll et al. 2001). It is noteworthy that many new explanations offered for the processes underlying moral decision-making employed the traditional explanatory pattern contrasting emotion and reason, with an emphasis on the former (Haidt 2007). However, it is important to emphasize the descriptive-normative divide, or the is-ought-problem, respectively, once again: Whereas the philosophers mentioned above debated which psychological faculty we should prioritize in order to make the morally right decision, most empirical researchers interpreted their evidence with respect to the faculties actually underlying the moral decisions their subjects made in a developmental context or an experiment. That is, they tried to explain how moral cognition and behavior works, not what is the morally right thing to do.

However, had the research been restricted to that descriptive question only, just offering an updated scientific account of how people of different kinds make moral decisions under certain circumstances, moral neuroscience might have just become a modernized form of moral psychology, promising to offer better explanations

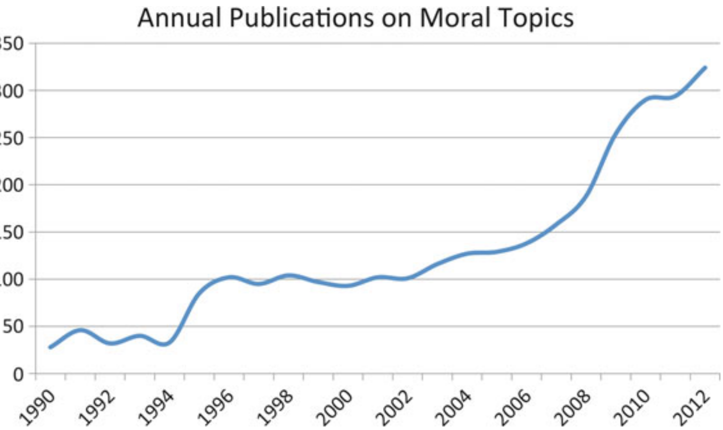

Fig. 7.1 The number of publications in the ISI Web of Science on moral topics was smaller than 50 until 1994, fluctuated around 100 until 2002, and increased steeply since then up to 324. The increase covers all three sub-databases for Social Sciences, Science and Technology, and Arts and Humanities publications. This pattern is not just due to the overall increase of annual publications in the Web of Science with 4.6 million in 2000 up to 7.3 million in 2012 (1.6-fold increase), because the number of publications on moral topics rose from 93 to 324 in the same period (3.5-fold). Method: Topic Search with at least one phrase match of "moral" in combination with any of the following items: judgment, emotion, reasoning, decision, cognition, behavior, and behavior. That is, every publication counted here contains at least one instance of, for example, the phrase "moral cognition" as its topic in the database

owing to its more direct access to the central organ of the mind, the human brain. The seductive allure, from the neuroscientific point of view, and the provocation, from the philosophical point of view, instead consisted in the attempt to cross the border between the descriptive and the normative, for example, by distinguishing morally justified ("rational") intuitions from unjustified ("irrational") ones, based on the brain areas, associated psychological processes, and evolutionary pathways putatively underlying them (Greene et al. 2004, 2008; Singer 2005). Some believed that in the perceived stalemate between different kinds of moral theories, where philosophers sometimes used imagined or actual moral dilemmas to undermine a theory by showing that the assumptions underlying it or decisions following from it are counterintuitive, neuroscience might lend a helping hand.

It probably were such attempts that invited others to critically investigate the explicit or implicit normative presumptions of such arguments (e.g., Kahane 2011), sometimes concluding that the attempts were normatively completely insignificant (e.g., Berker 2009) or, by contrast, identifying genuinely new ways in which moral psychology and neuroscience could and perhaps even should enrich moral philosophy (e.g., Joyce 2008; ▶ Chap. 10, "Psychology and the Aims of Normative Ethics," this section). It goes without saying that some philosophers took the occasion to fundamentally inquire into the meaning of essential concepts of the research such as "moral intuition" (e.g., ▶ Chap. 11, "Moral Intuition in Philosophy

and Psychology," this section). Concurrently, psychologists and neuroscientists questioned the empirical validity of some of the proposed explanations of moral cognition, sometimes re-analyzing or re-interpreting available data (e.g., McGuire et al. 2009; Moll and de Oliveira-Souza 2007), sometimes carrying out improved follow-up experiments (e.g., Moore et al. 2008). Investigating moral cognition has indeed become a diverse and fruitful endeavor attracting empirical as well as theoretical contributions from many disciplines (Waldmann et al. 2012). The following chapters of this section aim to add their share:

The Chapters in this Section

Chelsea Helion and David A. Pizarro, both at Cornell University in Ithaca, NY, USA, start out with a great service to the readers who are not very familiar with the state of the art in moral psychology: In their paper *Beyond dual-processes: The interplay of reason and emotion in moral judgment*, they review many essential developments in that discipline since the late 1990s, that is, the period when psychologists increasingly started to doubt that reason alone accounts for moral judgments. Moreover, experts in the field are likely to profit from engaging with Helion's and Pizarro's argument, stating that psychology's popular dual-process models (Kahneman 2011) and psychologically inspired philosophy do not adequately explain moral cognition. Dual-process models are models that usually distinguish quick, spontaneous reactions, often related to intuition and emotion, from slower, deliberative ones, often associated with reason and cognition.

Discussing the case of disgust in more detail, the authors emphasize that emotion and reason as well as their supposed interaction should not be understood too simplistically: After all, emotional processes are not simply something given that directly influence moral judgment. Rather they are themselves processes that can be regulated and adapted to different situations. Helion and Pizarro criticize that emotion regulation, a process studied intensively in other branches of psychology, has hitherto not been considered sufficiently in research on moral judgment. Depending on the moral context, they argue, people can upregulate or downregulate emotions based on their specific moral beliefs and goals, such as when a vegetarian intensifies his or her feelings of disgust in response to cruelty toward animals. This could explain why people experiencing the same emotions initially might arrive at different moral decisions.

While Helion's and Pizarro's proposal implies that we should beware of simplified psychological and philosophical understandings of emotion and reason and thus makes explanations more complex, it also promises a deeper understanding of moral cognition by drawing our attention not only to the particular circumstances and contexts in which people make moral decisions, but also to their further moral attitudes and beliefs that influence the way in which emotions affect judgments. In contrast to simple models that reduce moral decision-making only to emotion or only to reason and in contrast to dual-process models that take both kinds of processes into account but still presume a clear distinction between them,

Helion and Pizarro conclude that the cognitive cannot be separated strictly from the affective. Although parsimony or simplicity is often considered as virtue of scientific explanations, they cannot be virtuous if they are too reductive to do justice to the phenomena under investigation; it will be interesting to see whether an emphasis on emotion regulation and the complex association between emotion and reason has more explanatory power than other popular accounts within moral psychology and neuroscience.

In a similar way, DaniloBzdok and Simon Eickhoff at the Research Center Jülich together with DominikGroß at the University of Aachen, Germany, provide a service to readers unfamiliar with the scientific literature, but now with a focus on moral neuroscience. After first summarizing several key findings of years of neuroimaging research on moral decision-making, they raise the general question whether the many singular findings of brain areas activated during moral cognition found in previous studies can be integrated into a large-scale network. To answer it, they use a method called Coordinate-Based Meta-Analysis in combination with an activation likelihood estimation algorithm to test for consistent patterns of brain activation across those studies.

Bzdok, Eickhoff, and Groß discuss the results of their meta-analysis particularly with respect to three broad cognitive categories popular in social and cognitive neurosciences, namely, theory of mind, that is, thinking about somebody else's mental states, empathy, and mind-wandering, that is, lying in a brain scanner without any external stimulation. As they show in their paper *The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering*, activation in areas frequently associated with these cognitive categories is often also seen in studies of moral cognition, though to varying degrees. This shows, according to the authors, that there is no dedicated morality module in the human brain, or in their own words, that "no part of the brain is uniquely devoted to moral cognition but this capacity is very likely deployed across several heterogeneous functional domains."

While this finding, simplified to the conclusion that morality, when looking at the brain, is essentially social, may not be surprising to some readers, the authors discuss important experimental and theoretical implications of their results: The suggestion that moral cognition may not be a unified psychological entity raises the question whether there could actually be different kinds of moral cognition that might nevertheless be identified and/or separated in further studies. The diversity on the neuroscientific level could reflect the diversity of the concept of morality itself. Further, their outcome might reveal a general bias in experimental designs used to investigate moral decision-making, often relying on abstract, possibly not very ecologically valid stimuli restricted to moral cognition, not moral *behavior*. Bzdok, Groß, and Eickhoff conclude with a discussion of the meaning of their findings for neuroethics, particularly the descriptive-normative-divide or the is-ought-problem, respectively. They emphasize what we should be careful not to neglect it.

Their general finding is actually reminiscent of Helion's and Pizarro's argument not to underestimate the complexity of moral cognition: Just as it is unlikely that this can be accounted for by only a few basic psychological processes, it is unlikely

that it can be reduced to the brain activity of only a few basic areas. It is an interesting question for further analysis what the brain activity overlap between emotion regulation and moral cognition would look like, a research question involving the expertise of both groups of authors. Surprisingly though, the psychological as well as the neuroscientific review both invite us to reconsider basic questions of what moral cognition actually is and what is the best way to investigate it empirically, basic questions as they are particularly essential to philosophical inquiry, to which we now turn with respect to the next two chapters:

Regina Rini at the University of Oxford, UK, starts out with a service to those who are not very familiar with the scope and aims of moral philosophy: In her paper *Psychology and the Aims of Normative Ethics*, she poses the very general question what normative ethics actually is about, how it is defined. However, just as noted before that there is not only one understanding of "morality," she points out that it is unlikely that there is only one definition of "normative ethics" either. Rini nevertheless identifies three essential questions of that field: First, what a moral agent is; second, which actions are morally permitted or even required; and third, which moral beliefs are justified.

She tries to answer these questions in a dialectical fashion in order to show how moral philosophy can benefit from moral psychology and neuroscience. That is, she starts out with a negative answer why empirical research on moral cognition and decision-making is not relevant to each of the three questions and counters them with a positive rebuttal subsequently, discussing classical sources as well as recent contributions of normative ethics. Rini supports each of her positive answers with instructive examples. Generally, to show that moral psychology or neuroscience has at least some relevance to moral philosophy, it would suffice that she provides a convincing argument in only one of the three cases.

While Rini presents these arguments in favor of the normative significance of the empirical investigation of moral cognition, she concedes that the disciplinary interactions are still at a very early stage and concludes with a number of difficult but important questions for future research: Should it not matter if moral theories demanded something of us that we are psychologically incapable of? Or should this rather motivate us to change our psychological constitution? Would a better scientific understanding of the origins of our moral beliefs undermine their justification and ultimately lead to moral skepticism, the view that *none* of our moral beliefs are justified? In contrast to the humility of her eventual conclusion on psychology and the aims of normative ethics, Rini is quite convinced that psychologists – understood in a broad sense – and philosophers have a lot of work to do together.

Antti Kauppinen at Trinity College, Dublin, UK, contributes his share to that joint work for psychologists and philosophers, in relation to the problem of moral skepticism that Rini briefly refers to in her concluding paragraph. In his paper *Moral Intuition in Philosophy and Psychology*, he investigates whether and under which circumstances moral intuitions may have a justificatory role – but not without first clarifying many of the different possible definitions of "intuition" in philosophy and psychology, particularly when understood as a psychological state.

In a systematic manner, Kauppinen first analyzes different notions of "intuition" in contemporary empirical moral psychology and neuroscience, discussing the dual-process model, the relation between intuition and explanation in the sense that intuitions are understood as proximal causes of moral judgments, that is, that they explain to some extent why we make particular moral decisions, and research suggesting their putative unreliability. He continues, second, with a discussion of the role of intuitions in moral philosophy, namely, two kinds of Intuitionism and Rawlsian Coherentism, and eventually analyzes how intuitions could or could not have a role in justifying these accounts.

Although Kauppinen takes empirical research on moral cognition into account throughout his paper, he particularly devotes his third and last part to a reconciliation of the rather psychological and rather philosophical parts. After emphasizing once again that (moral) intuitions could be quite different things within the different frameworks, he critiques the dual-process model in a way that is reminiscent of Helion's and Pizarro's critique of its simplistic account of emotions, namely, by emphasizing that we should not take intuitions as something that is independent of someone's further beliefs. Indeed, Kauppinen subsequently clarifies the relation between the concepts of "emotion" and "intuition" in the end and summarizes his view regarding the circumstances under which intuitions can play "at least a quasi-foundational role in moral justification."

In the fifth and last paper of this section, titled *The half-life of the moral dilemma task – a case study in experimental (neuro-) philosophy*, Stephan Schleim at the University of Groningen, The Netherlands, carries out a deep analysis of the philosophical as well as the psychological questions that inspired the possibly most popular experiments in moral neuroscience, the moral dilemma task as investigated by Joshua Greene and collaborators (Greene et al. 2001, 2004). With what he calls "The Experimental Neurophilosophy Cycle" he particularly wants to inform those readers not very familiar with experimental research themselves of the translational procedures involved in designing, carrying out, and interpreting this kind of research. With his illustrated "Cycle," Schleim proposes a template for analysis that can ideally be applied to other cases of experimental (neuro-) philosophy as well and that also invites people to reflect on the relation between philosophy and empirical research more broadly.

Central to Greene and colleagues' investigation, Schleim argues, was the psychological puzzle posed by different response patterns to different kinds of moral dilemmas involving the sacrifice of a smaller number of people to save the lives of a larger number, a difference that was in the end explained by differences in emotional responses – reminiscent of what Kauppinen called "proximal causes" – and even used to undermine the justification of some kind of moral theories while supporting that of others, in particular utilitarianism. However, as Schleim argues, these interpretations presume a couple of intermediary steps, such as a choice of experimental procedures and methods, carrying out the actual experiment, preprocessing and analyzing data, and interpreting them in a psychologically significant way, making use of inferences to the best explanation. Just as the

strength of a chain depends on the strength of its individual links, the strength of the whole "Cycle" depends on the strength of its constitutive parts.

In his conclusion, Schleim notes that in light of his neurophilosophical analysis, philosophers need not fear to lose their jobs due to the explanatory power of cognitive neuroscience, as once was suggested in a report accompanying the original publication of the moral dilemma task study. By contrast, just as it is the case with respect to philosophy of science in general, philosophers and scientists alike can – and perhaps even should – engage in a critical philosophy of experimentation to emphasize the individual choices, experimental limitations, and backgrounds of explanatory patterns to better understand the scope of the research and in particular the strength of the translational links connecting philosophy, psychology, and neuroscience.

Possible Implications for Applied Ethics

From the perspective of neuroethics, the topic of this section is very abstract and theoretical and some may wonder whether it may have any implications for applied (neuro-) ethics at all. Although it is interesting in itself to understand how moral cognition, an essential human capacity, works and although the answers to this question may influence our understanding of what it means to be human, it is conceivable that the research might become relevant to a couple of applied questions, as the following examples are intended to show:

First, informed consent is an important issue in medical ethics and practice in general; it might be a particular challenge in the case of neurological or psychiatric/psychological disorders where the central decision-making structure may be impaired (Northoff 2006). Thus, finding out what the necessary and sufficient structures are for moral cognition may help to understand when a subject still is or is no longer capable of making an informed decision in a morally salient situation such as agreeing to undergo a potentially painful and risky medical treatment, perhaps even more so in cases where subjects are so much impaired that they cannot communicate any more in a normal manner, such as in a minimal conscious state (Jox et al. 2012). The issue of informed consent might be particularly problematic when the cognitive-emotional capacities required for it are precisely the target of the intervention and will only be sufficiently restored after the treatment.

Second, clinically as well as nonclinically, the bodily functions underlying moral cognition might themselves become a target of diagnosis and intervention. For example, in the history of psychiatry, some mental disorders, particularly psychopathy, had a moral connotation or were (and sometimes still are) understood as a moral disease (Werlinder 1978). The moral skills of psychopaths are a common topic of scholarly debate (Kennett and Fine 2008; Levy 2007; Sauer 2012), and it is very likely that there will be more research within moral psychology and neuroscience on that disorder. Already now, efforts are made to diagnose or identify psychopaths by means of brain scanners (Anderson and Kiehl 2012) and also to

screen and intervene (Rose 2010), that is, to treat them with the aim of curing their alleged moral deficits. It has even been discussed under which conditions research on psychopaths' psychological capacities might influence current legal practices of responsibility and legal insanity, emphasizing the emotional aspect prevalent in the moral neuroscience literature (Morse 2008). Besides the question whether this will ever be feasible, it goes without saying that the possibility of such interventions, affecting the deep core of someone's personality, needs critical discussion by ethicists, medical experts, and those primarily affected by the decisions alike.

However, even those who are not considered to be morally impaired might once be confronted with the option of intervening in the bodily functions underlying moral cognition, namely, third, as a form of moral enhancement (Douglas 2008). The previously mentioned moral dilemma task was actually already used in experiments to test whether subjects, when under the influence of a psychopharmacological substance like a selective serotonin re-uptake inhibitor, make "better" moral decisions than a control group treated with placebo (Crockett et al. 2010). It may be just as difficult as in the formerly mentioned case of moral treatment to find a means of moral enhancement that works, especially outside of a protected laboratory environment, and to agree on what a better moral decision would be in the first place. It goes without saying that any such attempts would have wide social ramifications.

Fourth and finally, it should not be forgotten that many moral psychologists, particularly those who carried out their research from a developmental point of view, developed ideas on improving moral education generally (Kohlberg 1984; Lind et al. 2010). If moral neuroscience truly adds to our understanding of moral decision-making, the developmental and situational conditions under which it functions, then we might expect this knowledge to contribute to the traditional pedagogical aim. This would not have to be restricted to pupils and students, but might even include preventive training to anticipate and avoid moral failure in governmental or private institutions.

As stated above, moral cognition as a research topic is abstract and theoretical, involving many conceptual as well as experimental challenges (Waldmann et al. 2012). However, this short and certainly preliminary list shows that there are related applied issues as well, some of which will probably increase in relevance as the field develops further. The last point for consideration mentioned here is the public as well as scholarly communication of the results, which often suggested that our moral decisions are made arbitrarily, that spontaneous emotional or intuitive responses may mislead us, even that we can be morally "dumfounded," that is, that we tend to cling to a judgment once made even after we learned that the reasons usually justifying this judgment do not apply in this particular instance. A popular case reported by Jonathan Haidt was an (imagined) example of incest that excluded all of the usual counterarguments, such as coercion, minority, or the increased risk of ill offspring, that people still considered as morally inappropriate after it had become clear that their justifying reasons do not apply in the presented case (Haidt et al. 1993).

However, other psychologists have warned that the widely disseminated findings of what John Kihlstrom called the "people are stupid" school of psychology,

presenting our decisions as emotionally driven, not based on reasoning, irrational, blinded by egoistic interests, and easily manipulated in experimental settings may constitute a limited and one-sided scientific account of what people are like (Kihlstrom 2004; Turiel 2010). The papers in this section emphasize that many central notions and experimental designs are preliminary; therefore, their results and interpretations are preliminary, too. Furthermore, much of the research is based on the subjects' trust in the experimenter, subjects who are often medicine, psychology, or neuroscience students and sometimes even participate in their own professors' experiments for mandatory course credit. The rational decision these subjects make is to actually participate in these experiments, in which they are often not informed or perhaps even misled about the experiments' aim – for otherwise the experimental manipulation might not work anymore. The behavior they subsequently show is also a function of the way the experiment is designed. The papers in this section suggest that there may be more than just one account to be told about what people are like, particular with respect to their moral cognition.

Acknowledgments I would like to thank Michael von Grundherr, Felix Schirmann, and Steffen Steinert for helpful suggestions to improve a previous version of this chapter.

Cross-References

▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
▶ Moral Intuition in Philosophy and Psychology
▶ Psychology and the Aims of Normative Ethics
▶ Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges
▶ Responsibility Enhancement and the Law of Negligence
▶ The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy
▶ The Morality of Moral Neuroenhancement
▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering
▶ The Use of Brain Interventions in Offender Rehabilitation Programs: Should It Be Mandatory, Voluntary, or Prohibited?

References

Anderson, N. E., & Kiehl, K. A. (2012). The psychopath magnetized: Insights from brain imaging. *Trends in Cognitive Sciences, 16*(1), 52–60.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs, 37*(4), 293–329.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences of the United States of America, 107*(40), 17433–17438.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*(3), 228–245.

Gazzaniga, M. S. (2005). *The ethical brain*. New York/Washington, DC: DANA Press.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nvstrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog. *Journal of Personality and Social Psychology, 65*(4), 613–628.

Hume, D. (1777/1975). *Enquiries concerning human understanding and concerning the principles of morals* (3rd ed.). Oxford: Oxford University Press.

Jox, R. J., Bernat, J. L., Laureys, S., & Racine, E. (2012). Disorders of consciousness: Responding to requests for novel diagnostic and therapeutic interventions. *Lancet Neurology, 11*(8), 732–738.

Joyce, R. (2008). What neuroscience can (and cannot) contribute to metaethics. In W. Sinnott-Armstrong (Ed.), *The neuroscience of morality: Emotion, brain disorders, and development* (Vol. 3, pp. 371–394). Cambridge, MA: MIT Press.

Kahane, G. (2011). Evolutionary debunking arguments. *Noûs, 45*(1), 103–125.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). New York: Farrar, Straus and Giroux.

Kant, I. (1785/2011). *Groundwork of the metaphysics of morals: A German-English edition* (trans: Gregor, M., & Timmermann, J.). Cambridge: Cambridge University Press.

Kennett, J., & Fine, C. (2008). Internalism and the evidence from psychopaths and "acquired sociopaths". In W. Sinnott-Armstrong (Ed.), *The neuroscience of morality: Emotion, brain disorders, and development* (pp. 173–190). Cambridge, MA: MIT Press.

Kihlstrom, J. F. (2004). Is there a "People are Stupid" school in social psychology? *Behavioral and Brain Sciences, 27*(3), 348-+.

Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages* (1st ed.). San Francisco: Harper & Row.

Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, & Psychology, 14*(2), 129–138.

Lind, G., Hartmann, H. A., & Wakenhut, R. (2010). *Moral judgments and social education*. New Brunswick: Transaction Publishers.

Littlefield, M. M., & Johnson, J. M. (2012). *The neuroscientificturn: Transdisciplinarity in the age of the brain*. Ann Arbor: University of Michigan Press.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology, 45*(3), 577–580.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences, 11*(8), 319–321.

Moll, J., Eslinger, P. J., & de Oliveira-Souza, R. (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task – Preliminary functional MRI results in normal subjects. *Arquivos De Neuro-Psiquiatria, 59*(3B), 657–664.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*(6), 549–557.

Morse, S. J. (2008). Psychopathy and criminal responsibility. *Neuroethics, 1*(3), 205–212.

Nadelhoffer, T., Nahmias, E. A., & Nichols, S. (2010). *Moral psychology: Historical and contemporary readings*. Malden: Wiley-Blackwell.

Northoff, G. (2006). Neuroscience of decision making and informed consent: An investigation in neuroethics. *Journal of Medical Ethics, 32*(2), 70–73.

Piaget, J. (1932). *Le jugement moral chez l'enfant*. Paris: LibrairieFèlix Alcan.

Pickersgill, M., & van Keulen, I. (Eds.). (2012). *Sociological reflections on the neurosciences*. Bingley: Emerald.

Pizarro, D. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour, 30*(4), 355-+.

Rose, N. (2010). 'Screen and intervene': Governing risky brains. *History of the Human Sciences, 23*(1), 79–105.

Sauer, H. (2012). Psychopaths and filthy desks. *Ethical Theory and Moral Practice, 15*(1), 95–115.

Sayre-McCord, G. (2012). Metaethics. In *The Stanford encyclopedia of philosophy* (Spring 2012 Ed.), from http://plato.stanford.edu/archives/spr2012/entries/metaethics/

Singer, P. (2005). Ethics and intuitions. *Journal of Ethics, 9*, 331–352.

Turiel, E. (2010). Snap judgment? Not so fast: Thought, reasoning, and choice as psychological realities. *Human Development, 53*(3), 105–109.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. Oxford: Oxford University Press.

Werlinder, H. (1978). *Psychopathy: A history of the concepts: Analysis of the origin and development of a family of concepts in psychopathology*. Uppsala/Stockholm: Almqvist &Wiksell International.

Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment

8

Chelsea Helion and David A. Pizarro

Contents

Abstract

A great deal of research in moral psychology has focused on the interplay between emotion and reason during moral judgment, characterizing the two as forces working in opposition to influence judgment. Below, recent psychological research on morality is reviewed, with a special focus on disgust and the nature of its role in moral and political judgment. Behavioral, neuroscience, and physiological data are reviewed looking at the role of disgust in moral judgment, with a particular emphasis on the role of emotion regulation – the process of shifting emotional responses in order to meet one's goals. It is suggested that dual-process theories of moral judgment are not well suited to understand the role of emotion regulation in influencing moral judgments and decisions. Theories that emphasize the primacy of one process over another may ultimately be missing the complexity how these processes interact to influence moral judgment.

C. Helion (✉) • D.A. Pizarro
Department of Psychology, Cornell University, Ithaca, NY, USA
e-mail: cah298@cornell.edu; pizarro@gmail.com

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_160,

Introduction

> It was on the moral side, and in my own person, that I learned to recognize the thorough and primitive duality of man; I saw that, of the two natures that contended in the field of my own consciousness, even if I could rightly be said to be either, it was only because I was radically both
> -Robert Louis Stevenson 1886, p. 56

Stevenson's classic 1886 novel "The Strange Case of Dr. Jekyll and Mr. Hyde," tells the tale of Dr. Jekyll, a man who wants to end the struggle between good and evil within him. In order to do so, he formulates and ingests a potion designed to split his mind into two – one personality to house his baselessness and immorality (Mr. Hyde), and one to house the morally pure traits he values most. The inner conflict that humans experience between their moral selves and their more unrestrained, egoistic selves has been a consistent theme in literature for centuries. While (largely) discarding the good-versus-evil aspects of this dichotomy, moral psychology has nonetheless embraced the basic division of mental processes into two general types – one mental system that is cold, rational, and deliberative, and another that is emotional, intuitive, and quick. This characterization has served as a basic organizational framework for understanding the processes involved in human judgment across a variety of domains, including moral and ethical judgments (e.g., Kahneman 2011). This "dual-process" approach to the mind has motivated a great deal of research on moral judgment within the last decade, and has led psychologists to reconsider the historically dominant approaches to moral judgment, approaches which emphasized the primacy of reason. But the division of the mind into two systems, while fruitful, has encouraged researchers to characterize emotion and reason as forces acting upon moral judgment in opposing directions, and to focus on factors that give rise to the dominance of one over the other. This chapter will focus on a particular emotion – disgust – in order to illustrate that the simplicity of the dual-process approach may hide some of the more nuanced ways in which emotion and reason interact to produce moral judgment and behavior. In particular, it will highlight research on emotional regulation as an example in which reason guides emotion rather than battles against it. Taken together, these two bodies of research suggest that characterizing reason and emotion as separate, opposing forces in moral judgment is a caricatured description of two processes that often interact in complex ways to motivate judgment and action (Pizarro 2000).

The Dethroning of Reason

For the majority of the twentieth century, psychologists viewed decision making as a "cold" process in which individuals calmly and rationally weigh pros and cons to arrive at a decision that maximizes their utility (Loewenstein and Lerner 2003). Research in moral psychology echoed this emphasis on reason: Early studies of

moral judgment[1] focused on the importance of cognitive abilities across development, and on how the emergence of specific abilities shaped the child's understanding of moral rules (Kohlberg 1963; Piaget 1932). As individuals matured, they would approach a higher "stage" of moral thinking, and while many would never reach the highest stages of moral reasoning, the majority of individuals would reach a stage of moral sophistication required to uphold the norms and laws of a functional society.

However, as research on emotion grew, it began to transform the behavioral sciences, and the study of judgment and decision making in particular. No longer was judgment characterized as a "cold" emotionless process, but as infused with emotion at every turn. In an influential paper, Haidt (2001) built upon this emerging affective science, and applied it to moral judgment by making a radical claim – that reason played a much smaller role in ethical and moral judgment than psychologists thought. Taking a note from the philosopher David Hume (1777/1960), Haidt argued that intuitions (often in the form of emotional responses) were, in fact, the primary causes of moral judgment. In what he dubbed "social intuitionism," Haidt proposed that when individuals are faced with a moral question (e.g., "Julie and Mark are a brother and sister that had consensual sex"), and are asked to make a moral judgment ("How morally wrong was Julie and Mark's behavior?"), it is their experience of emotion (in this example, disgust) that gives rise to the moral judgment ("What Mark and Julie did was wrong!"), not their ability to formulate reasons. Notably, in examples like this, individuals are actually at a loss to justify their moral judgment and become "morally dumbfounded" – reduced to offering something like "it is just gross" as their only reason for their judgment.

For Haidt, ethical judgment is a great deal like aesthetic judgment: It is made quickly, effortlessly, and without a great deal of conscious deliberation (Greene and Haidt 2002). More radically, Haidt claimed that if deliberate reasoning played a role at all in moral judgment, it was most likely as a post hoc attempt to justify an individual's intuition-driven moral judgments (Haidt 2001; Haidt and Bjorklund 2008). As support for his claim, Haidt also drew on growing evidence from social psychology, demonstrating the power of nonconscious influences on judgment, as well as from cross-cultural research showing that cognitive-developmental theories (such as that of Kohlberg) did a poor job at predicting moral judgments in non-Western cultures (Haidt et al. 1993; Shweder et al. 1997).

Haidt's social intuitionist model happened to coincide with a new approach to study morality pioneered by Greene and his colleagues, who paired modern neuroimaging techniques with classic philosophical moral dilemmas to arrive at a similar conclusion – that emotions played a much larger role in moral judgment than previously thought. Greene and colleagues (2001) demonstrated that when faced with classic sacrificial moral dilemmas (in which one individual must be

[1]Notably, some early researchers suggested that children's morality was largely emotionally driven (Wendorf 2001; Kline 1903). The insights from this approach were all but lost in the coming dominance of the cognitive-developmental approach.

sacrificed to save a greater number of people), individuals often responded in a manner consistent with their gut, emotional reactions, and argued that this could be seen in the patterns of neural activation observed while individuals were reading emotionally evocative dilemmas. However, when given enough time to deliberate, individuals could overturn their initial gut reactions and reply with a more calculating (i.e., utilitarian) response. Yet for both Greene and Haidt, the emphasis was on the divided mind – one system that produces an initial moral judgment, and another capable of intervening occasionally to alter the judgment. While these approaches were influential in shifting moral psychology away from the historically dominant rationalist theories' of moral judgment, the pendulum swung swiftly in the direction of emotional primacy – the view that reasoning played little or no role in most moral judgments. Indeed, combined with the research emerging in other areas of judgment and decision making, it seemed fairly obvious that human judgment was driven if not solely, at least primarily, by emotional and nonrational processes (Bargh and Ferguson 2000).

Yet a number of more recent findings have demonstrated that reason is not as inert as these accounts (or, at least, the stronger versions of these accounts) implied. For instance, there is evidence that simply temporarily encouraging rational thought can have a demonstrable effect on moral judgment. Researchers recently demonstrated that engaging in a cognitively challenging task prior to making a moral judgment can cause individuals to go against their initial intuitive moral judgment when responding to moral dilemmas (Paxton et al. 2011). In addition, when given sufficient time to deliberate, individuals are more likely to be persuaded by reasoned arguments that a seemingly wrong, yet harmless act was not, in fact, immoral (Paxton et al. 2011).

Encouraging individuals to adopt a rational mindset can lead them to make judgments of moral blame that go against their intuitive initial reaction. For instance, individuals intuitively reduce their judgments of blame for harmful acts that were intended, although the causal link between the intention and the harm occurred in an unexpected fashion (Pizarro et al. 2003). However, when instructed to think rationally and deliberately, individuals become insensitive to extraneous information about the causal link and focus on the intention of the actor and the harmful outcome. In other words, a simple prompt to respond rationally is enough to shift the nature of moral judgment toward answers that are more careful, deliberate, and free of normatively extraneous information.

One final source of evidence for the ability of reason to influence moral judgment comes from research demonstrating that the likelihood of engaging in moral reasoning changes based on cognitive resources available to the individual at the time the moral judgment is made. Occupying an individual's mind with a task (i.e., placing them under "cognitive load") has been shown to interfere with the moral reasoning process, with individuals under load more likely to favor intuitive moral decisions (Greene et al. 2008). Cognitive load can also derail the moral justification process, reducing the likelihood that individuals will rationalize their own immoral behavior (Valdesolo and DeSteno 2008). The corollary of this evidence is that, in the absence of constraints on the ability to deliberate, individuals are indeed making

use of careful deliberation when arriving at moral judgment. In short, there is plenty of evidence that individuals can (and do) engage in moral reasoning when motivated or prompted to do so.

Yet even research emphasizing the influence of reason adheres to a fairly simplistic dichotomy pitting reason against emotion/intuition. This may be because researchers tend to utilize methods designed to pit one process against the other, or that favor one process over the other (Monin et al. 2007). The answers to emerge from these methods will, by design, seem as evidence for or against one side of the dichotomy. For instance, Kohlberg's (1963) methodology involved presenting participants with moral dilemmas that pit two equivalent courses of action against each other, and to ask participants to verbally justify their decisions. When presented with the famous "Heinz" dilemma (in which participants must determine if it is better for Heinz to steal an expensive medicine in order to keep his wife from dying), participants are not only asked to determine which is the better decision, they are also asked detailed questions about their reasoning process. In its explicit prompting of reason, and in its use of dilemmas with equally compelling courses of action, any researcher would conclude that reasoning is at the heart of moral judgment (Monin et al. 2007).

On the other hand, researchers who favor emotional/intuitive accounts of moral judgment often ask participants to evaluate the moral infractions committed by others, frequently focusing on extreme scenarios involving bestiality, child pornography, and incest (for reviews see Pizarro and Bloom 2003; Monin et al. 2007). Unlike the Kohlbergian dilemmas, these moral scenarios create a strong, immediate affective reaction (and are not "dilemmas" in the traditional sense of the word, since one course of action seems clearly worse than another). Faced with such scenarios, participants rarely need to deliberate about competing moral principles in order to arrive at a judgment. Moreover, rather than being asked to reason aloud to justify their moral judgments, participants are generally asked to assess moral wrongness after the putative moral infraction has been committed (rather than to debate the possibilities for a future course of action). These kinds of reaction-focused questions tend to stack the deck in favor of an emotion-based account of moral judgment.

Despite these methodological limitations, however, there is still a great deal of evidence pointing to a more complex interrelation between reason and emotion/intuition. Perhaps shifting the question from simply asking if reason influences moral judgment, and toward *when* and *how* reasoning influences moral judgment yields more nuanced insight. Take one example of a more subtle interaction between reason and intuition – studies of expertise have shown that a learned process can be made intuitive over time (Zajonc and Sales 1966). Similarly, moral intuitions themselves may arise from prior instances of moral reasoning. A person may reason their way to a particular moral view (such as that animals should not be killed and eaten, or that slavery is wrong), and over time, this moral view becomes intuitive (Pizarro and Bloom 2003). The mechanism by which a reasoned choice may become an intuition is, unfortunately, not well-captured in a dual-process approach that does not take into account the ways in which intuition and reason interact over time.

Disgust and Moral Judgment: A Brief Overview

When it comes to the regulation of human behavior, many moral codes extend far past the concerns over harm and justice that have traditionally been the focus of the cognitive-developmental tradition in moral judgment. A growing body of work has demonstrated that moral codes emphasize a number of other domains – respect for authority, group loyalty, and purity. For instance, large sections of the Bible, Koran, and many other religious texts focus on the importance of keeping oneself clean in body and in spirit (Haidt et al. 1993; Shweder et al. 1997). The motivational fuel that enforces moral norms across these domains appears to be emotions such as anger, empathy, contempt, guilt, shame, gratitude, and (of particular relevance to this discussion) disgust (Rozin et al. 1999; Decety and Batson 2009; Lazarus 1991; Trivers 1971).

Disgust – an emotion that likely evolved to prevent individuals from coming into contact with dangerous pathogens – has recently been strongly linked to moral, social, and political judgments (Schnall et al. 2008; Inbar et al. 2009a, b). A great deal of research has been conducted in the past two decades in an attempt to classify when, where, and to whom disgust is expressed (Haidt et al. 1994; Rozin et al. 2000; Olatunji et al. 2008; Tybur et al. 2009). It has become clear that disgust is strong, easy to induce, and provides immediate motivation to avoid the target of disgust. There also appear to be a set of near-universal elicitors – bodily and animal products such as feces, urine, vomit, and rotten meat (all potential transmitters of pathogens) seem to elicit disgust in most people. In addition, often all it takes is a single picture or word to make an individual feel full-blown disgust. In this sense, it is one of the least "cognitive" emotions – it can often seem more like a reflex.

Research on individual differences in disgust has also shed light on the nature of disgust elicitors. The most widely used measure of individual differences in disgust is the Disgust Sensitivity Scale (DS-R; Haidt et al. 1994, modified by Olatunji et al 2007). This scale divides disgust into three unique sets of elicitors: core, animal reminder, and contamination disgust. A more recent scale proposes a different set of subdomains of disgust: pathogen disgust, sexual disgust, and moral disgust (Tybur et al. 2009). The authors suggest that each facet of disgust fits a specific evolutionary function: Pathogen disgust (analogous to core/contamination disgust) is meant to protect an individual from disease, sexual disgust is meant to protect an individual from actions that would stand in the way of one's evolutionary fitness (e.g., incest, individuals that one does not find aesthetically or histologically attractive), and moral disgust is meant to protect an individual from those that would hurt the success of the individual or the group (such as acts of selfishness). Nonetheless, both scales emphasize the fact that disgust appears to be a response to potential contamination from a substance, individual, or action.

While disgust may not be the most relevant moral emotion, nor even the most common, disgust is focused on because the wealth of research on this emotion helps shed light on a more general point – that the interaction between reason and emotion in moral judgment is far more complex than one might expect (Pizarro et al. 2011). A great deal of evidence accumulated in the last decade suggests that

disgust easily extends its influence to the sociomoral domain – individuals use disgust terminology to describe both the revolting and the reviled (Rozin et al. 2008). While it may have originated as an emotion meant to keep us from ingesting something dangerous, it now seems to motivate us to keep away from individuals or entire social groups, and to evaluate a certain kind of act as morally wrong. For instance, feeling disgust at the time of moral judgment increases the severity of a moral judgment – people think that an act is more wrong, and that an individual is more blameworthy – even when the source of disgust is completely unrelated to the target of judgment (Schnall et al. 2008). While Schnall and colleagues (2008) found a domain-general increase in moral severity, recent research has shown that feeling disgust may play an especially strong role in moral judgments having to do with *purity*, acts seen as wrong not because of their direct harm but because of their symbolic degradation or contamination of the self or society. Feeling disgust in response to purity violations (such as consensual incest) has been linked to more severe punishment for moral transgressors (Haidt and Hersh 2001).

Indeed, disgust is especially powerful in influencing judgments in the domain of sexual mores. Inbar, Pizarro, Knobe, and Bloom (2009b) found that people who are more easily disgusted (as measured by the "disgust sensitivity" scale; Olatunji et al. 2007) have more negative implicit attitudes about homosexuality. Echoing this, individuals who are easily disgusted are more likely to show negativity toward homosexuals, but not toward other out-groups such as African-Americans (Inbar et al. 2012b; Tapias et al. 2007). In addition, inducing disgust increases people's explicit and implicit bias against homosexuals (Dasgupta et al. 2009; Inbar et al. 2012b).

Individual differences in the tendency to feel disgust (i.e., *disgust sensitivity*) has also been linked more generally to political conservatism, specifically in political issues related to purity and divinity – such as abortion and gay marriage (Inbar et al. 2009a, 2012a). This is consistent with work showing that liberals and conservatives rely upon different kinds of moral intuitions for their judgments – liberals rely on assessments of harm and fairness when making moral judgments, whereas conservatives also rely on purity, loyalty, and authority (Graham et al. 2009). Indeed, simply reminding people that there is disease in the environment (consistent with the motivation induced by disgust) can lead individuals to temporarily report being more politically conservative (Helzer and Pizarro 2011).

One of the more interesting ways in which disgust has been implicated in moral judgment comes from work on what some researchers have dubbed "moralization." Within a generation (and perhaps not even that), we can observe concrete changes in societal views concerning the morality of certain acts. For instance, while in the 1960s, smoking was ubiquitous, today, smoking is confined to select areas, smokers are shown the door and asked to partake elsewhere, and morally judged for their behavior (Rozin 1999). How did a behavior like smoking – which was so commonplace 50 years ago – become moralized over time? Rozin (1999) implicates disgust in this process of moralization – bringing a behavior that was previously seen as non-moral into the moral domain. Rozin and Singh (1999) showed that the targets

of moralizing disgust can even change across one's life span. They surveyed college students, their parents, and grandparents, and found that all three groups reported being equally as disgusted by and expressive of negative attitudes toward cigarette smoking, even though the grandparents indicated that they had grown up in an age that was more tolerant toward cigarette smoking.

Researchers are increasingly learning about the neural and physiological correlates of disgust. Experiencing and recognizing disgust has been linked to activation in the anterior insula and putamen (Moll et al. 2002; Calder et al. 2000); however, this relationship is not consistently found across all disgust studies that utilize neuroimaging techniques (Phan et al. 2002). Disgust has also been associated with greater facial muscle tension, both increased and decreased heart rates, and increased skin conductance (Demaree et al. 2006). Olatunji and colleagues (2008) found differences in the physiological reactions between different kinds of disgust: Core and contamination disgust (such as the disgust over rotten meat, or at sipping out of a stranger's beverage by mistake) were associated with increased facial muscle tension and heart rate while watching a video of an individual vomiting, and watching a video and having blood drawn, was associated with higher facial muscle tension and *decreased* heart rate in individuals sensitive to "animal reminder" disgust (disgust related to gore, body-envelope violations, and dead bodies).

Emotion Regulation: The Intersection of Reason and Emotion

Knowing how disgust works to influence moral, social, and political judgments is informative, but it paints an incomplete picture of how emotions (like disgust) influence individuals over time. A key limitation of many studies on emotion is that they do not take into account the various strategies individuals employ in everyday life to either avoid feeling certain emotions, feel them less strongly, or feel them more strongly. In fact, it is fairly evident that individuals engage in this sort of emotional regulation fairly frequently (Gross 2002). This regulation is necessary, in part, because the environment in which emotions evolved is in many ways quite dissimilar to the current environment, making emotional responses in the modern world poor guides to achieving goals (Tooby and Cosmides 1990; Gross 1998a). This ability to regulate emotions allows, more generally, for a rich interaction between an individual's long-term, deeply valued goals and her short-term emotional reactions. In the case of moral judgment, the need for emotional regulation should be clear – individuals often need to alter their emotional states to coincide with their moral goals.

Researchers investigating the regulation of emotion have proposed five different categories of emotional regulation (Ochsner and Gross 2008): (1) *situation selection* – selecting situations that are conducive to attaining one's goals or to avoid ones that are not (for example, a married man declining an invitation to grab a drink with an ex-girlfriend), (2) *situation modification* – taking steps to alter one's current situation to bring it in line with one's goals (if the man does accept the invitation,

choosing to talk about his happy marriage instead of reminiscing about the past relationship), (3) *attentional deployment* – focusing one's attention on something else (choosing to focus on how gray his ex-girlfriend's hair has become rather than on her ample cleavage), (4) *cognitive change* – changing one's emotional understanding of the situation at hand by cognitively reappraising features of the situation (reframing the situation as catching up with an old friend rather than drinking with a former lover), and (5) *response modulation* – regulating the physiological response of an emotional state while it is currently being experienced (the man telling himself that his sweaty palms are due to the crowded bar rather than to any feelings of attraction). The first four components of emotional regulation have been referred to as *antecedent-focused* regulation strategies, and the fifth is referred to as a *response-focused* regulation strategy (Gross 1998b).

Previous research has indicated that regulating negative emotions, and specifically disgust, can have downstream cognitive and physiological consequences. Multiple studies have asked participants to adopt an antecedent-focused (e.g., reappraisal) or response-focused (e.g., suppression) regulation strategy, and have demonstrated that each makes different contributions to altering one's emotional experience. Gross (1998b) had participants watch a disgust-eliciting video, and found that though both reappraisal and suppression reduced disgust-expressive behavior, reappraisal decreased ratings of subjective disgust while suppression had no effect on subjective disgust, and was instead linked to increased activation of the cardiovascular system. Recent research has demonstrated that this type of reappraisal process can be automated via the use of implementation intentions – regulatory strategies that take the form of an if-then plan – and that different implementation intentions can affect what aspect of the disgust experience is regulated (Schweiger Gallo et al. 2009; Gallo et al. 2012). Gallo and colleagues (2012) had participants form either a goal intention ("I will not get disgusted!"), an antecedent-focused implementation intention ("I will not get disgusted, and if I see blood, I will take the perspective of a physician!"), or a response-focused implementation intention ("I will not get disgusted, and if I see blood, I will stay calm and relaxed!") before reporting on valence and arousal while viewing a series of disgusting and non-disgusting images. They found that individuals who had formed an antecedent-focused implementation intention reported that the disgusting images were significantly less negative, but that there were no differences between this group and the goal-intention group on reported arousal, suggesting that this antecedent-focused strategy was changing the meaning of the emotional experience without altering the physical experience. Individuals who had formed a response-focused implementation intention reported significantly less arousal when viewing disgusting images as compared to the other two groups; however, there were no differences between this group and the goal-intention group on assessments of valence. Taken together, these studies suggest that different emotion regulation strategies can alter different components of the emotional experience. Within the moral domain, it remains unclear what aspects of disgust experience (valence, arousal, appraisal) working alone or in tandem contribute to moral judgment, and using different antecedent- or response-focused strategies to regulate disgust may help illuminate this process.

The Neuroscience of Emotion Regulation

We now know a great deal more about the neural underpinnings of emotion regulation, and how emotion and reason interact within the brain. For instance, a study that used functional neuroimaging to look at different regulatory strategies showed that emotional suppression and reappraisal work on different time courses, specifically showing that when asked to regulate disgust, reappraisal was linked to increased activation in prefrontal areas (the medial and left ventrolateral prefrontal cortex – areas associated with cognitive control) during early stimulus presentation, and was correlated with decreased activity in regions known to be implicated in affective responses (left amygdala and insula) during the later stages of stimulus presentation. Emotional suppression showed a distinctly different pattern, and was linked to activation of prefrontal control areas during the later stages of stimulus presentation, accompanied by increased amygdala and insula responses (Goldin et al. 2008). This suggests that different regulatory strategies may play out over a different time course, and that they have a differential impact on the subjective, physiological, and neural components of an emotional experience.

A great deal of research has implicated the prefrontal cortex (PFC), a region associated with volition, abstract reasoning, and planning, as playing a primary role in the process of emotion regulation (Ochsner and Gross 2005; Wager et al. 2008). Emotion regulation appears to engage multiple areas of the PFC, including the dorsolateral prefrontal cortex (dlPFC), the ventrolateral prefrontal cortex (vlPFC), and the dorsomedial prefrontal cortex (dmPFC) (Ochsner and Gross 2005). In addition, research suggests that successfully reappraising emotional stimuli involves both cortical and subcortical pathways, roughly illustrating that the process recruits areas of the brain associated with "cognitive" and "affective" processes (Wager et al. 2008). For instance, the amygdala, a subcortical structure heavily implicated in affective responses, plays an integral role in the processes of guiding attention to and forming evaluations of affective stimuli (Ochsner 2004). The amygdala's detection of affective stimuli can happen rapidly and can even occur non-consciously (Amodio and Devine 2006). Further supporting its role as a key player in emotional regulation, amygdala activation and deactivation has been linked to the augmentation and reduction (respectively) of an affective response (Ochsner and Gross 2005). Other affective brain regions involved in emotional regulation include the ventral striatum, the mid-portion of the cingulate cortex, and the insula – an area that has been implicated in the subjective experience of disgust and has been of particular importance in linking disgust with moral judgment (Lieberman 2010).

A greater understanding of the interactions between affective brain regions and higher cognitive brain regions during emotional regulation may help shed light on both the psychology of regulatory behavior and on an understanding of how emotion regulation may inform moral judgment. A great deal of research in emotion regulation and neuroimaging has focused on cognitive reappraisal, an antecedent-focused regulation strategy that involves reframing emotionally evocative events. Many of these studies involve presenting participants with aversive

images during a functional MRI scan, while giving them instructions on how to view the image in ways that may encourage the up- or down-regulation of their emotional response. Using this method, Wager and colleagues (2008) demonstrated that cognitive reappraisal relies on a bidirectional relationship between affective and cognitive regions. They found that the cognitive reappraisal of emotion involves the successful recruitment of areas associated with memory, negative affect, and positive affect/reward. Specifically, they found that the relationship between the left vlPFC (an area involved in higher cognition) and reappraisal success involves two mediating pathways: (1) a path which predicts reduced reappraisal success involving areas involved in negative emotion, such as the amygdala, lateral orbitofrontal cortex (OFC), and anterior insula, and (2) a path predicting increased reappraisal success involving areas implicated in positive affect/reward, such as the ventral striatum, the pre-supplementary motor area (SMA), the precuneus, and subgenual and retrosplenial cingulate cortices. The positive association between left vlPFC activation and the activation of both of these networks suggests that the left vlPFC plays a role in both the generation of (path 1) and regulation of (path 2) negative affect during cognitive reappraisal.

In short, the ability to successfully regulate emotion relies on structures implicated in the generation of both negative and positive affect, as well as on the same structures being able to both reduce negative appraisals and generate positive ones. What this suggests is that the regulatory strategy of cognitive reappraisal has properties that overlap significantly with both systems – affective and cognitive. This echoes a claim made by Pizarro and Bloom (2003), who pointed to the importance of cognitive appraisals in guiding moral responses that are typically described as emotional. Taken together, this research suggests that the emotional reactions that accompany a moral evaluation can be regulated via cognitive reappraisal, allowing for a great deal of flexibility on the influence that emotions (like disgust) play in the formation of moral judgments.

This bidirectional relationship between emotion and cognition makes sense within the context of moral judgment. In the classic trolley dilemma, an individual is asked to imagine that they are a trolley car operator and that a runaway trolley is hurtling down the track (Foot 1967; Thomson 1985). The trolley has to go on one of two diverging tracks: (1) a track where five people are working and (2) a track where one person is working. In a typical moral psychology experiment, participants are then asked about the permissibility of killing one to save five. Using this dilemma, Greene and colleagues (2001) uncovered the role of emotional engagement in moral judgment. To manipulate emotional engagement, participants were presented with two versions of trolley-style dilemmas: (1) In the *impersonal* version, participants are told that they can hit a switch that will put the trolley onto a different track, where it will only hit one person, (2) In the *personal* version, participants are asked to imagine that they are standing next to a large stranger on a footbridge that goes over the trolley tracks, and if they push the stranger, the trolley will stop, thus saving the five people.

The researchers found that increased emotional engagement (personal vs. impersonal) elicited greater activation in regions of the brain that had been implicated in affective processing (the bilateral medial frontal gyrus, the bilateral posterior cingulate gyrus, and the left and right angular gyrus). In the impersonal moral condition, they observed significantly more activation in regions associated with working memory (the right middle frontal gyrus, the left parietal lobe, and the right parietal lobe). Greene and colleagues (2004) extended this result, and showed that participants exhibited greater activation in the amygdala when they are resolving personal moral dilemmas than when they resolving impersonal moral dilemmas. During the personal moral dilemmas, participants also exhibited increased activation in brain regions implicated in theory of mind processes: the medial prefrontal cortex (mPFC), the precuneus, the posterior superior temporal sulcus (pSTS), and the temporoparietal junction (TPJ). The researchers used this as evidence to make the claim that personal moral dilemmas are more affectively charged, and further suggest that personal moral dilemmas involve a network that focuses attention away from the "here and now" and instead directs attention to predicting future events and considering the mental states of others (Greene 2009). More recent work has modified the impersonal/personal distinction, instead focusing on psychological and neural differences between deontological judgments (which some have posited are automatic responses driven by emotion) and utilitarian or consequentialist judgments that some claim are the product of conscious moral reasoning (Greene 2009; Greene et al. 2008). Nonetheless, the revised distinction retains the distinction between "emotional" and "cognitive" processing that gives rise to different kinds of moral judgments.

Though the existence of two distinct systems is a plausible account for the observed pattern of results, reconciling this work with research in emotion regulation perhaps prompts a slightly different description regarding the processes involved in guiding these sorts of judgments. Rather than characterizing judgment as driven by two opposing processes fighting over which answer is morally correct, these dilemmas are prompting individuals to reconcile their affective responses with their moral goals through the regulation of their emotional reactions. Though we tend to think of the typical instance of emotional regulation as the down-regulation of an emotional response, there are times when individuals up-regulate their affective responses in order to meet their goals. Within the moral domain, this is particularly the case for empathy, where taking the perspective of another is often accompanied by increased emotional arousal for the self (Batson 1998). The personal and impersonal versions of the trolley dilemma may just as easily be described as involving cognitive appraisals that facilitate the up- and down-regulation of emotional experiences, and that those who are able to regulate their emotions effectively are able to suppress or increase the affective response that they view as appropriate for the dilemma at hand.

One source of evidence for the importance of such up- and down-regulatory strategies comes from research demonstrating that manipulating the self-relevance of emotional stimuli (akin to the personal/impersonal distinction in the moral research) can influence one's affective experience (Ochsner et al. 2004). In one

study, participants were asked to up- or down-regulate their emotions using a self-focused reappraisal strategy (i.e., to think about the personal relevance of each image as it appeared). For example, if participants were shown a picture of a gruesome car accident, participants were asked to either imagine themselves or a loved one in the negative situation (up-regulation) or to think of the situation from a detached third-person perspective (down-regulation). Participants reported that down-regulating emotion was significantly more difficult than up-regulating emotion. In addition, amygdala activation was modulated by reappraisal, with up-regulation being linked to increased activation in the left amygdala and down-regulation was linked to bilateral amygdala deactivation.

This self-focused reappraisal strategy may be analogous to the personal/impersonal moral distinction, in which individuals are asked to put themselves in the situation of physically pushing a man to his death or to physically distance themselves from the event by imagining themselves flipping a switch. Asking participants to imagine themselves causing or being personally involved in a situation is similar to a self-focused up-regulatory strategy, whereas asking participants to imagine pushing the button is a self-focused down-regulatory strategy. Thus, it seems possible that the differences observed in the personal/impersonal moral dilemmas may reflect the effects of up- and down-emotion regulation, rather than the workings of two distinct processes. This would be consistent with the suggestion that individuals who favor utilitarian solutions to affectively charged sacrificial dilemmas are either simply less likely to feel a negative affective reaction to the dilemma in the first place, or are able to down-regulate their negative emotional reactions in order to meet their utilitarian moral goals (Bartels and Pizarro 2011).

Conclusion and Future Directions

The most plausible approach regarding the processes involved in emotion regulation, especially, we think, in the domain of moral judgment, is what researchers have termed an "instrumental" account of regulation. This account breaks from the tradition of straightforward psychological hedonism – the view that individuals are always motivated to feel positive emotions and minimize negative emotions – and instead suggests that emotion selection and regulation can be described as maximizing the utility of a particular goal, even if the goal is best served by feeling a negative emotion (Tamir 2009). Certain emotions may be more useful in some contexts than others – while pleasant emotions may be selected for when immediate benefits are greater than long-term benefits (e.g., smiling when one's child presents them with a homemade drawing), when long-term benefits are greater, individuals may instead want to feel a helpful emotion, or one that will help them meet long-term goals (e.g., expressing anger when said drawing has been scrawled in permanent marker on the living room wall).

Applying this framework to disgust, it seems possible that individuals may encourage their emotions of disgust when evaluating particular moral acts or

individuals in order to effectively communicate disapproval and rejection of immoral behaviors (Hutcherson and Gross 2011). Gaining a better understanding of when individuals feel disgust within moral contexts and how this response relates to the individual's long- and short-term goals may help us understand the role that individual differences in the tendency to experience certain emotions (such as disgust sensitivity) play in forming moral judgments. For example, individuals may up-regulate their disgust within moral contexts (e.g., when a vegetarian intensifies their disgust in order to fuel their moral indignation about animal cruelty) or down-regulate (e.g., when a liberal reappraises two men kissing as an act of love) based on the current context of the judgment, and on their specific moral beliefs and goals.

The fact that two individuals who experience strong disgust arrive at different moral judgments makes more sense when taking into account the ability to regulate emotional responses, rather than assuming a static, linear relationship between emotion and judgment. This may be true of other emotional reactions as well: Individuals may up-regulate their anger when they are making judgments about punishment or assigning moral blame. It seems likely that one of the contributing factors to moral judgment is the ability to up- and down-regulate emotion, depending on the context of the moral situation and the cognitive and motivational resources that are available to the individual at the time of moral judgment – something that simple dual-process theories do not accommodate well. Yet the growing body of research looking at emotional regulation – which we believe should play a larger role in our psychological theories of morality – suggests that emotion and cognition are best viewed as a set of processes that are so deeply intertwined that it cannot be captured within a simple dichotomy. Individuals, using a variety of strategies, are able to selectively dampen or heighten emotional experiences – often in the service of their higher-order goals – and thus shape the contribution of their emotions to their moral judgments. In the same way that Jekyll cannot divorce himself from Hyde, human beings cannot divorce the cognitive from the affective. It appears that they are, quite literally, formed of the same stuffs.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Free Will and Experimental Philosophy: An Intervention
▶ Free Will, Agency, and the Cultural, Reflexive Brain
▶ Moral Cognition: Introduction
▶ Moral Intuition in Philosophy and Psychology
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play
▶ The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy
▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661.

Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin, 126*, 925.

Bartels, D., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154–161.

Batson, C. D. (1998). Altruism and prosocial behavior. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (Vol. 2, pp. 282–316). Boston: McGraw-Hill.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Reviews Neuroscience, 3*, 1077–1078.

Dasgupta, N., DeSteno, D. A., Williams, L., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion, 9*, 585–591.

Decety, J., & Batson, C. D. (2009). Empathy and morality: Integrating social and neuroscience approaches. In J. Braeckman, J. Verplaetse, & J. De Schrijver (Eds.), *The moral brain*. Berlin: Springer.

Demaree, H. A., Schmeichel, B. J., Robinson, J. L., Pu, J., Everhart, D. E., & Berntson, G. G. (2006). Up- and down-regulating facial disgust: Affective, vagal, sympathetic, and respiratory consequences. *Biological Psychology, 71*, 90–99.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.

Gallo, I. S., McCulloch, K. C., & Gollwitzer, P. M. (2012). Differential effects of various types of implementation intentions on the regulation of disgust. *Social Cognition, 30*(1), 1–17.

Goldin, P. R., McRae, K., Ramel, W., & Gross, J. J. (2008). The neural bases of emotion regulation: Reappraisal and suppression of negative emotion. *Biological Psychiatry, 63*(6), 577.

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives use different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences IV*. Cambridge, MA: MIT Press.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Gross, J. J. (1998a). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*, 271–299.

Gross, J. J. (1998b). Antecedent-and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology, 74*(1), 224.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39*, 281–291.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (The cognitive science of morality: Intuition and diversity, Vol. 2, pp. 181–217). Cambridge, MA: MIT Press.

Haidt, J., & Hersh, M. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31*, 191–221.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.

Haidt, J., McCauley, C., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences, 16*(5), 701–713.

Helzer, E., & Pizarro, D. A. (2011). Dirty liberals!: Reminders of cleanliness promote conservative political and moral attitudes. *Psychological Science, 22*, 517–522.

Hume, D. (1960). *An enquiry concerning the principles of morals*. La Salle, IL: Open Court. (Original work published 1777).

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*, 719–737.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009a). Conservatives are more easily disgusted. *Cognition & Emotion, 23*, 714–725.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009b). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*, 435–439.

Inbar, Y., Pizarro, D. A., Ayer, R., & Haidt, J. (2012a). Disgust sensitivity, political conservatism, and voting. *Social Psychological and Personality Science, 3*, 527–544.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012b). Disgusting smells cause decreased liking of gay men. *Emotion, 12*(1), 23.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kline, L. W. (1903). A study in juvenile ethics. *The Pedagogical Seminary, 10*(2), 239–266.

Kohlberg, L. (1963). Moral development and identification. In H. Stevenson (Ed.), *Child psychology: 62nd yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 143–193). New York: McGraw-Hill.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 619–642). New York: Oxford University Press.

Moll, J., de Oliveira - Souza, R., Eslinger, P. J., Bramati, I. E., Mourao - Miranda, J., & Andreiuolo, P. A. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience, 22*, 2730–2736.

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology, 11*(2), 99–111.

Ochsner, K. N. (2004). Current directions in social cognitive neuroscience. *Current Opinion in Neurobiology, 14*, 254–258.

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences, 9*, 242–249.

Ochsner, K. N., & Gross, J. J. (2008). Cognitive emotion regulation: Insights from social cognitive and affective neuroscience. *Current Directions in Psychological Science, 17*, 153–158.

Ochsner, K. N., Ray, R. D., Robertson, E. R., Cooper, J. C., Chopra, S., Gabrieli, J. D. E., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion. *NeuroImage, 23*(2), 483–499.

Olatunji, B. O., Williams, N. L., Tolin, D. F., Sawchuck, C. N., Abramowitz, J. S., Lohr, J. M., et al. (2007). The disgust scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment, 19*, 281–297.

Olatunji, B. O., Haidt, J., McKay, D., & David, B. (2008). Core, animal reminder, and contamination disgust: Three kinds of disgust with distinct personality, behavioral, physiological, and clinical correlates. *Journal of Research in Personality, 42*, 1243–1259.

Paxton, J. M., Ungar, L., & Greene, J. D. (2011). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage, 16*(2), 331.
Piaget, J. (1932). *The moral judgment of the child*. New York: Harcourt, Brace Jovanovich.
Pizarro, D. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour, 30*, 355–375.
Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review, 110*, 193–196.
Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39*, 653–660.
Pizarro, D., Inbar, Y., & Helion, C. (2011). On disgust and moral judgment. *Emotion Review, 3*(3), 267–268.
Rozin, P. (1999). The process of moralization. *Psychological Science, 10*, 218–221.
Rozin, P., & Singh, L. (1999). The moralization of cigarette smoking in America. *Journal of Consumer Behavior, 8*, 321–337.
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The moral-emotion triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral ethics (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76*, 574–586.
Rozin, P., Haidt, J., & McCauley, C. R. (2000). Disgust. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (2nd ed., pp. 637–653). NewYork: Guilford Press.
Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust: The body and soul emotion in the 21st century. In D. McKay & O. Olatunji (Eds.), *Disgust and its disorders* (pp. 9–29). Washington, DC: American Psychological Association.
Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096–1109.
Schweiger Gallo, I., Keil, A., McCulloch, K. C., Rockstroh, B., & Gollwitzer, P. M. (2009). Strategic automation of emotion regulation. *Journal of Personality and Social Psychology, 96*, 11–31.
Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity), and the "big three" explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.
Stevenson, R. L. (1886/2006). *The strange case of Dr. Jekyll and Mr. Hyde*. New York: Scribner.
Tamir, M. (2009). What do people want to feel and why? Pleasure and utility in emotion regulation. *Current Directions in Psychological Science, 18*(2), 101–105.
Tapias, M., Glaser, J., Vasquez, K. V., Keltner, D., & Wickens, T. (2007). Emotion and prejudice: Specific emotions toward outgroups. *Group Processes and Inter Group Relations, 10*, 27–41.
Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *279*, 1395–1415.
Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology, 11*, 375–424.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Tybur, J. M., Lieberman, D. L., & Griskevicius, V. G. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology, 29*, 103–122.
Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Psychology, 44*(5), 1334–1338.
Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M., & Ochsner, K. N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron, 59*, 1–14.
Wendorf, C. W. (2001). History of American morality research, 1894–1932. *History of Psychology, 4*(3), 272–288.
Zajonc, R. B., & Sales, S. M. (1966). Social facilitation of dominant and subordinate responses. *Journal of Experimental Social Psychology, 2*(2), 160–168.

The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering

9

Danilo Bzdok, Dominik Groß, and Simon B. Eickhoff

Contents

Abstract

A sense of morality forms the fabric of human societies. There is an ongoing debate whether the cognitive and emotional sources of moral decisions might be closely related to theory of mind, an abstract–cognitive capacity, and empathy, an automatic–affective capacity. That is, moral decisions are believed to imply representation of other individuals' thoughts and emotional states, respectively. Moreover, it has been noticed that neural activation patterns during moral cognition are very similar to the brain areas engaged during mind-wandering, i.e., neural correlates of an endogenously controlled state in the absence of a specific mental task.

Investigation of the neural substrates underlying moral cognition was greatly facilitated by the advent of neuroimaging techniques. This growing number of observation on brain activation patterns during the aforementioned tasks now

D. Bzdok • S.B. Eickhoff (✉)
Institut für Neurowissenschaften und Medizin (INM-1), Jülich, Germany
e-mail: danilo.bzdok@rwth-aachen.de; s.eickhoff@fz-juelich.de

D. Groß
Medical School, RWTH Aachen University, Aachen, Germany
e-mail: dgross@ukaachen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_161,

provides rich substrates for a quantitative integration of the current literature. Such large-scale integration, identifying brain areas consistently engaged by moral, social, empathic, and unconstrained cognition, then provides a quantitative basis for the comparison of their neuronal implementation. This chapter thus quantitatively assesses and reviews the neurobiological relationship between the moral network and the neural networks subserving theory of mind, empathy, and unconstrained cognition.

In conclusion, the neural network subserving moral decisions probably reflects functional integration of distributed heterogeneous networks, is dissociable into cognitive and affective components, as well as highly similar to the brain's default activity pattern.

Introduction

Moral behavior has classically been thought to be based on rational (i.e., rather conscious, controlled, and effortful) thinking. Rational explanations assumed that moral behavior arises from a conscious weighing of different rules, norms, and situational factors. In contrast, the role of emotion and intuition in moral thinking (thought to represent an unconscious, automatic, and effortless way) has been less often emphasized (Haidt 2001). Emotional explanations emphasized the influence of intuitive, subconscious emotional states that are rapidly evoked by a given situation. Taken together, *abstract–inferential* and *automatic–emotional* processing have been implicated and contrasted in philosophical, psychological, and biological accounts of moral behavior.

The association of psychological categories, such as decision-making or emotional influences thereon, with brain activity in the underlying neural networks has been greatly promoted by the development of functional neuroimaging. Positron emission tomography (PET) and the noninvasive functional magnetic resonance imaging (fMRI) allow the in vivo investigation of functional specialization in the human brain. Based on local changes in cerebral blood flow and glucose or oxygen metabolism, these techniques allow inference on regional increases in neural activation during the performance of specific tasks. Often, the neural correlates of a given task (reflecting a mental process of interest, e.g., moral decision-making) are isolated by subtraction of the activation measured during a closely related task (a control task, such as semantic or abstract decisions) that is supposed to carry the same confounds (e.g., reading) but not to evoke the mental process of interest. Over the last two decades, functional neuroimaging has then provided a wealth of information on the cerebral localization of various psychological tasks, including moral decision-making.

Notions of rationality and emotionality also serve as explanations in contemporary imaging research on the neural correlates underlying moral decisions (moral cognition). Joshua Greene (in the USA) and Jorge Moll (in Brazil) can probably be considered the protagonists in the ensuing debate. Results from fMRI studies by Greene and colleagues (Greene et al. 2001; 2004) were consistently interpreted as revealing a neuroanatomical dissociation between emotional responses and

subsequent explicit cognitive modulations in moral cognition. However, fMRI findings by Moll and colleagues (Moll et al. 2005a, 2006; Moll and Schulkin 2009) were interpreted as revealing various different psychological processes without any specific neural correlates, including group-oriented (i.e., pro-social) and self-oriented (i.e., egoistic) affective drives, in moral cognition.

It is important to note that the rational and emotional facets of moral cognition are, by theoretical arguments and empirical research, closely related to two other aspects of social interaction: theory of mind (ToM) and empathy. ToM refers to the ability to contemplate other's thoughts, desires, intentions, and behavioral dispositions by *abstract inference* (Frith and Frith 2003; Premack and Woodruff 1978). Evidently, moral decisions are influenced by whether or not an agent's action is perceived as intentional or accidental, which crucially relies on mental state inference, i.e., ToM. Consistently, behavioral data from subjects with high-functioning autism, known for impoverished ToM abilities, suggested an involvement of ToM in moral judgments, given that these individuals relied less on the agent's intentions and more on action outcomes (Moran et al. 2011). Empathy, on the other hand, refers to *intuitively* adopting somebody's *emotional state* while maintaining the self–other distinction (Decety and Jackson 2004; Singer and Lamm 2009). More specifically, empathy can be subdivided into (partially intertwined) emotional and cognitive components (Shamay-Tsoory et al. 2009). Phylogenetically and ontogenetically earlier "emotional empathy" is closely related to emotion contagion and simulation systems, while later developing, more advanced "cognitive empathy" is more related to perspective-taking and imagination systems. In particular, empathy is different from and tends to precede sympathy, which does not necessarily result in identical affect in the observer, but for instance in pity or compassionate love. In moral decisions, experiencing empathy was shown to alleviate harmful actions towards others (Eisenberger 2000). Conversely, deficient empathy skills are a clinical hallmark of psychopathic subjects and are believed to contribute to their morally inappropriate behavior (Hare 2003). Taken together, the aforementioned debate on the contribution of cognitive and emotional factors to moral decision-making may be reframed as the question whether the neural correlates of moral decisions are closer related to those of ToM or empathy – or whether there is a distinct moral module serving moral cognition.

The Neural Architecture of Moral Cognition

In the present analysis we tried to avoid the pitfalls of descriptive verbal summaries of neuroimaging results, which are inevitably subjective and hence potentially biased. Such critical verbal analyses tend to focus on a limited number of preselected aspects and tend to be biased by the authors' own adherence to a specific research area. In contrast to classical review articles, coordinate-based meta-analysis (CBMA) is hypothesis-free, data-driven, and, hence, objective by algorithmically weighing all results equally. As the CBMA method is not skewed by subjectivity, it precludes overinterpretation of expected, easily interpretable

findings and neglect of unexpected, barely reconcilable findings in neuroimaging research. CBMA might therefore help to point out consistent, yet frequently ignored findings (Eickhoff and Bzdok 2012).

Rather than a critical verbal discussion, we therefore recently performed a quantitative CBMA of the neuroscientific literature on moral cognition using the activation likelihood estimation (ALE) algorithm (Eickhoff et al. 2009). This approach represents a powerful approach to gain a synoptic and, in particular, unbiased view of distributed neuroimaging findings. In this context, ALE addresses the following key question: Where in the brain is the convergence of the included experiments' activation foci higher than expected by chance? ALE thus offers a unique opportunity to quantitatively measure concordance between neuroimaging results without the implicit assumptions of neuroanatomical terminologies, which are at times inconsistently employed. It also allows relating different neural processes to each other by performing meta-analyses on different neuropsychological phenomena. This integration and synthesis of neuroimaging data thus permits statistically defensible inference on the neural basis of psychological processes across a large number of experimental implementations and subjects samples.

The presented summary of neuroimaging studies on moral cognition included all those experiments that required participants to make (covert or overt) appropriateness judgments on actions of one individual towards others. In these studies, participants evaluated mainly textual, sometimes pictorial social scenarios with moral violations or dilemmas. While this approach has by far dominated the neuroscientific investigation of moral decisions, it should be noted, however, that it largely equates to assessing the neural correlates of identifying and judging moral violations. In contrast, they are less focused on "rightful action," i.e., the implementation of moral thoughts and knowledge in one's own behavior. Nevertheless, the analysis presented in Fig. 9.1 represents the current state of neuroimaging evidence for moral cognition. The obtained pattern of converged brain activation is in very good agreement with descriptive reviews of fMRI studies on moral cognition (J. Greene and Haidt 2002; Moll et al. 2005b). In the following, we will discuss the presumed functional implications of the individual brain areas that resulted as significant loci of convergence.

The *medial prefrontal cortex* (mPFC) is a highly associative brain region implicated in a range of complex tasks, such as action monitoring, free thought, autobiographical memory recall, and the perception of others. In fact, consistent activity in the mPFC during moral cognition was found all along its dorsoventral axis (simply put: from the upper to the lower parts of the middle front side of the brain), including the dorsomedial prefrontal cortex (dmPFC), frontopolar cortex (FP), and ventromedial PFC (vmPFC). From a conceptual perspective, the more dorsal versus more ventral parts of the mPFC are discussed to relate to cognitive versus affective processes, controlled versus automatic processes, explicit versus implicit social cognition, goal versus outcome pathways, as well as other-focus versus self-focus. Direct evidence for such fundamental distinction is however still limited. It is noteworthy that mPFC damage early in life can leave intellectual abilities intact while leading to hindered acquisition of social conventions and

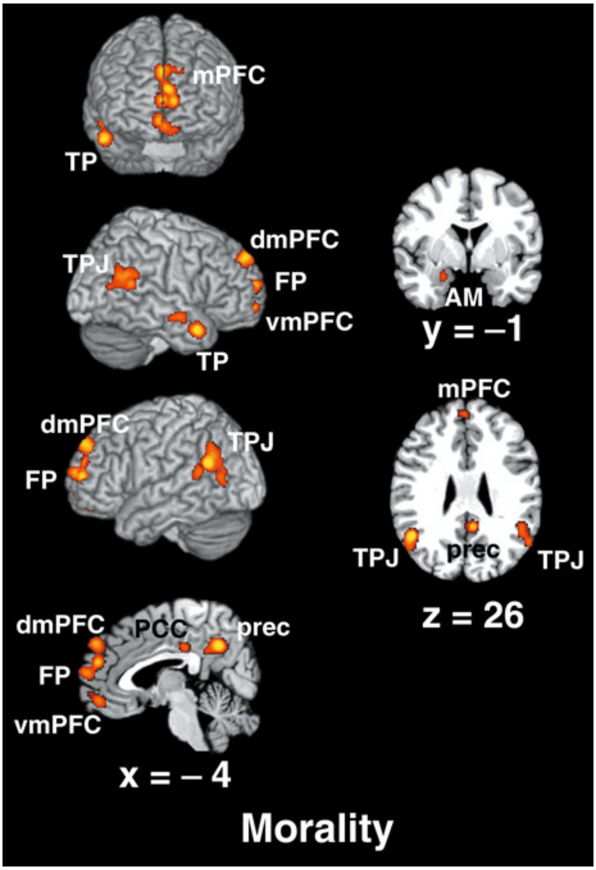

Fig. 9.1 Meta-analysis results on moral cognition. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to moral cognition (67 neuroimaging experiments). Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *AM* amygdala, *dmPFC* dorsomedial prefrontal cortex, *FP* frontal pole, *mPFC* medial prefrontal cortex, *PCC* posterior cingulate cortex, *prec* precuneus, *TP* temporal pole, *TPJ* temporo-parietal junction, *vmPFC* ventromedial prefrontal cortex (cf. Bzdok et al. 2012)

moral rules (Anderson et al. 1999). Early lesioned patients (much more than adult-onset prefrontal patients) display immoral behaviors, such as stealing, physical violence, and absence of remorse in the context of impaired moral reasoning (Moll et al. 2003). In short, a child's moral development can be disrupted by early mPFC damage.

The dmPFC has axonal connections with the temporo-parietal junction (TPJ; especially connected to superior frontal gyrus) and the precuneus (especially connected BA8/9), which have both likewise been implicated in the meta-analysis of moral cognition. The *temporo-parietal junction* is a supramodal association area whose heterogeneous functional profile seems to range from attentional reallocation, filtering irrelevant stimuli, and prediction generation over processing embodied self and predicting others' action intentions to agency. Paralleling its functional diversity, the TPJ literature offers various neuroanatomical labels for this area, including the terms angular gyrus, inferior parietal lobule, posterior superior temporal sulcus, supramarginal gyrus, BA 39, PGa/PGp, as well as "pli courbe." Consequently, interpretation of the inconsistently named TPJ can be challenging

given increased metabolic activity across disparate psychological tasks, stimulus domains, and experimental modalities.

The *precuneus* is another highly integrative area, which is believed to generate internally directed thoughts in form of self-referential visuospatial imagery (Cavanna and Trimble 2006). Consistently, the precuneus appears to mediate covert reallocation of spatial attention, that is, spatial cognition in the absence of physical (e.g., eye) movements (Gitelman et al. 1999), which led to its informal nickname "mind's eye." This proposed domain-spanning role might potentially explain its various domain-specific functional involvements, such as in visual rotation, deductive reasoning, autobiographical memory retrieval, and mental navigation in space.

The *posterior cingulate cortex* (PCC) is adjacent to but distinct from the precuneus by its connections to the limbic system and thus close relation to emotion processing (Margulies et al. 2009). This area is most frequently proposed to be important for the modality-independent retrieval of autobiographical memories and their integration with current emotional states (Maddock 1999).

As another affect-related brain region, the *amygdala* (AM) is believed to automatically extract biological significance from the environment and to shape appropriate behavioral responses (Bzdok et al. 2012; Sander et al. 2003). This functional concept covers its involvement in classical conditioning, emotion regulation, social cognition, reward processing, and memory formation (Adolphs 2010; Bzdok et al. 2012; LeDoux 2000). Considering that the amygdala is probably the brain area most unequivocally linked to emotion processing and given its heightened activity across the meta-analyzed neuroimaging studies, the reverse inference on an involvement of emotional brain systems in moral cognition seems justified.

Finally, the *temporal pole* (TP; here liberally referring to the entire anterior temporal lobe/BA38) was repeatedly proposed to store verbal and nonverbal semantic knowledge, in particular, context-independent social semantic knowledge, including values and concepts of social events (cf. Olson et al. 2007; Ross and Olson 2010; Zahn et al. 2007). Examples of such conceptual social knowledge would be the meaning and ramifications of "deceitfulness" or how people dress appropriately according to given situations. In line with this, neurological lesion of the TP entails social semantic deficits, such as failing to name human actions or to recognize the name, voice, handwriting, odors, or face of familiar people. Moreover, they may result in behavioral and personality disturbances, ranging from compulsively eating flower decorations on tables to general apathy to other's distress (cf. Gorno-Tempini et al. 2004).

It can be concluded that moral cognition is neurally implemented by brain areas that tend to be highly interconnected, not specific for any single psychological task, and not dependent on a specific sensory modality but rather multimodal and "associative." Those brain areas are moreover implicated in complex psychological processes, including social cognition, autobiographical memory retrieval, mental imagery, and reallocation of attention to internal information, all of which might contribute to the final "psychological outcome," i.e., moral judgment.

Two brain areas that have also been repeatedly discussed to subserve moral cognition, however, were not revealed by the meta-analysis. First, the posterior

superior temporal sulcus (pSTS) is involved in audiovisual integration and processing biologically significant movement (Hein and Knight 2008) and was reported to increase neural activity during moral cognition in several papers. The lack of convergence in the pSTS may readily be explained by inconsistent neuroanatomical labeling. As a rule of thumb, activation in the surroundings of this cortical area was often interpreted as "pSTS" in the morality literature and as "TPJ" in the ToM literature. The convergent activation in the meta-analysis on moral cognition is, however, neuroanatomically corresponding to the TPJ. That is, observed recruitment of the TPJ during moral tasks was perhaps recurrently mislabeled as pSTS, which might have confused discussion of the TPJ and pSTS in previous neuroimaging studies on moral cognition.

Second, the delineated "moral network" did not include the dorsolateral prefrontal cortex (dlPFC), conventionally interpreted as important for attention shifting and decision-making, although brain activity in this area was discussed in several original papers and reviews in moral neuroscience. The dlPFC was specifically proposed to reflect the engagement of abstract reasoning processes and cognitive control in moral cognition (Greene et al. 2004). Heightened dlPFC activity was thus argued to promote utilitarian responses by overriding prepotent socio-emotional behavioral tendencies. The absence of *consistent* metabolic increase in the dlPFC during moral decisions might be parsimoniously explained by selective recruitment. That is, this brain region might be recruited by the "core" moral network as an auxiliary functional module depending on the specific cognitive set imposed by specific moral decisions. Rather than being part of the "core" network, the dlPFC might have been observed to increase activity in difficult personal moral judgments and approving personal moral violations because of those decisions' increased cognitive demand. The same principle of context-dependent recruitment of supplementary areas might hold true for other brain regions that were associated with moral cognition repeatedly but not consistently, including but not restricted to the anterior insula, anterior cingulate cortex, and lateral orbitofrontal cortex. Also for these regions, inconsistent neuroanatomical labeling cannot be excluded as a confounding factor in the previous literature on moral cognition.

The Neurobiological Relationship Between Moral Cognition and Theory of Mind

Performing another meta-analysis on brain activity evoked by theory of mind then allowed elucidating the correspondence of the neural substrates consistently engaged by this task and moral cognition, two psychologically related mental processes (cf. introduction). We included those neuroimaging studies into the meta-analysis of theory of mind that required participants to adopt an intentional stance towards others, that is, predict their thoughts, intentions, and future actions. These studies mostly presented cartoons and short narratives that necessitated understanding the beliefs of the acting characters. The results of the meta-analysis

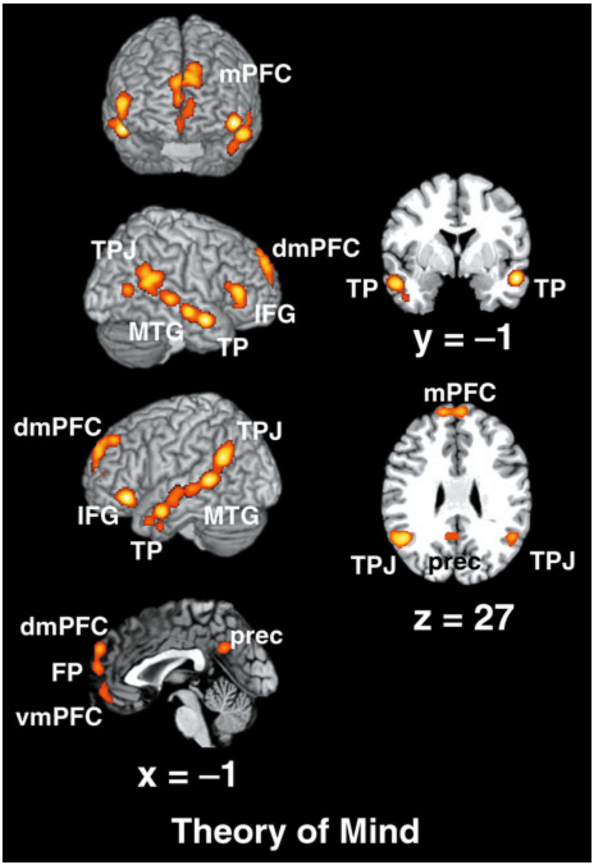

Fig. 9.2 Meta-analysis results on theory of mind. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to theory of mind (68 neuroimaging experiments). Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *dmPFC* dorsomedial prefrontal cortex, *FP* frontal pole, *IFG* inferior frontal gyrus, *MTG* middle temporal gyrus, *mPFC* medial prefrontal cortex, *prec* precuneus, *TP* temporal pole, *TPJ* temporo-parietal junction, *vmPFC* ventromedial prefrontal cortex (cf. Bzdok et al. 2012)

of ToM (Fig. 9.2) are consistent with earlier meta-analysis of such neuroimaging studies (Spreng et al. 2009). Conceptually, the convergence across studies on ToM resulted in an *abstract–inferential social–cognitive network* implicated in the recognition and processing of others' *mental states*.

Brain activity patterns during moral cognition and ToM overlapped in the bilateral vmPFC, FP, dmPFC, and TPJ, as well as the right TP (Fig. 9.3). This extensive convergence indicates that moral cognition and ToM engage a highly similar neural network. The homologous neural implementation, in turn, entices to speculate about a close relationship between these two psychological processes and the experimental tasks to probe these. The interest in the neurobiological relationships between moral cognition and ToM recently gained momentum, which entailed publication of a small number of targeted neuroimaging studies.

A seminal fMRI study investigated the interaction of a protagonist's initial intention and subsequent action outcome by explicit moral judgments of short written stories (Young et al. 2007). The bilateral TPJ, dmPFC, and precuneus

Fig. 9.3 Relationship between the moral network and the neural network underlying theory of mind. Overlapping activation patterns between the meta-analysis on moral cognition and the difference analysis between ToM and empathy. Coordinates in MNI space (cf. Bzdok et al. 2012)

showed significant signal effects for the interaction of negative versus neutral beliefs versus outcomes. The right TPJ showed the biggest signal increase in attempted (intention) but failed (outcome) harm, that is, when nothing bad actually happened despite what the protagonist planned. Given that moral cognition and mental state attribution were probably part of the participants' cognitive set in all experimental conditions, right TPJ activity appeared to reflect special emphasis on the agent's thoughts when weighing various contextual features against each other. Consistently, transient disruption of right TPJ activity using repetitive transcranial magnetic stimulation was found not to impair moral judgments per se (Young et al. 2010). Rather, this manipulation reduced the influence of the protagonist's belief without completely eliminating it as an input to judgment formation. It is important to note, however, that further evidence from neuroimaging (Mitchell 2008) and lesion (Apperly et al. 2007) studies questioned the *specificity* of the TPJ for belief processing. Nevertheless, right RTPJ activity appears to be, comparing to other relevant areas, particularly related to processing mind states in explicit moral cognition.

The ensuing notion that the TPJ may represent a crucial link between moral cognition and ToM was confirmed by another fMRI study that set out to detail encoding and integration of intentions in the context of moral judgments (Young and Saxe 2008). While "encoding" consists in merely creating a representation of the protagonist's belief, "integration" then consists in flexibly weighing the moral judgment depending on the interaction of intention and outcome. The bilateral TPJ and precuneus were related to both encoding the protagonist's belief and integrating it with other relevant contextual features. In fact, brain activity in these regions did not differ according to belief content, in particular, its valence (negative vs. neutral). In contrast, the dmPFC was related to processing belief valence during the integration phase. The authors thus proposed that the TPJ and precuneus mainly process beliefs, while the dmPFC mainly processes morally relevant aspects of the presented stories in constructing a coherent moral judgment. Analogous to the TPJ, it is important to note that dmPFC activity might not be *specific* to belief processing as it was for instance also linked to language coherence (Ferstl and von Cramon 2002). In addition to belief, outcome, and valence, metabolic responses in the dmPFC and bilateral TPJ during moral judgments were observed to vary according to the previously experienced fairness of the judged agent (Kliemann et al. 2008). This suggests that the dmPFC and bilateral TPJ might also integrate available memory of an agent's personality traits in explicit moral judgments. Another fMRI study showed selective metabolic increase in the dmPFC, right TPJ, and precuneus in response to morally relevant facts in short stories without explicit mental state information or an explicit necessity for moral judgments (Young and Saxe 2009). This finding can be taken to argue that brain regions typically related to ToM might be implicated not only in *explicit or controlled* but also *implicit or automatic* moral cognition.

In line with the meta-analytic overlap, the reviewed fMRI studies suggest that moral cognition might involve reconstructing personality attributes and intentions

of agents as well as their integration with action outcomes and other relevant contextual features when reasoning about morally significant social scenarios. It should, however, not be underestimated that experimental similarities (i.e., employed stimuli and paradigm) between studies on moral cognition and ToM might have contributed to the observed neural homology. In other words, the congruency may appear larger than it actually is due to shared low-level features (textual descriptions, cartoon, requirement to make a judgment).

The Neurobiological Relationship Between Moral Cognition and Empathy

In contrast to ToM, the correspondence between moral cognition and empathy has hardly been investigated in targeted neuroimaging research despite their relatedness on psychological grounds (cf. introduction). The correspondence between the neural substrates reported in those largely separate lines of research was therefore juxtaposed by individual meta-analysis on either topic (Figs. 9.1 and 9.4). We included those neuroimaging studies into the meta-analysis of empathy that aimed at eliciting the conscious and isomorphic (i.e., happiness in other induces happiness in oneself) experience of somebody else's affective state. Put differently, in these studies participants were supposed to "feel into" and thus know what another person was *feeling* (rather than *thinking*, which would be related to ToM). These studies employed mostly visual, sometimes textual, or auditory stimuli that conveyed affect-laden social situations which participants watched passively or evaluated on various dimensions. Conceptually, the convergence across studies on empathy resulted in an *automatic–emotional social–cognitive network* implicated in vicariously mapping others' *affective states*.

Please note that a meta-analytic distinction between emotional and cognitive empathy cannot and should not be done at this point. It cannot be done because there are currently not enough available neuroimaging studies on cognitive empathy. It should not be done because assuming a clear-cut neurobiological dissociation between emotional and cognitive empathy would constitute a fairly strong a priori hypothesis about how psychological constructs map on brain organization.

Brain activity related to both moral cognition and empathy converged significantly in an area of the dmPFC (Fig. 9.5). An fMRI study identified a similar area as highly selective for processing guilt (Wagner et al. 2011), an emotion closely related to moral and social norm violation. More specifically, guilt was proposed to promote interpersonal relationships by immediately providing actual or anticipated emotional feedback for the acceptability of actions (Tangney et al. 2007). Moreover, the dmPFC has consistently been related to the (possibly interwoven) reflection of own and simulation of others' mind states (Bzdok et al. 2013). One might therefore cautiously conclude that convergence in this highly associative cortical area might reflect complex representational social–emotional processing. Additionally, the individual meta-analyses revealed the left AM in moral cognition

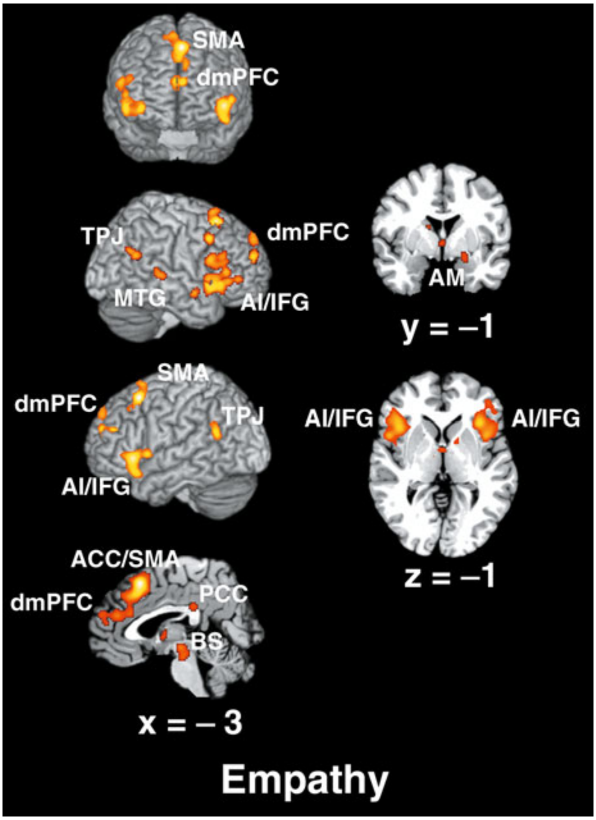

Fig. 9.4 Meta-analysis results on empathy. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to empathy (112 neuroimaging experiments). Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *ACC* anterior cingulate cortex, *AI* anterior insula, *AM* amygdala, *BS* brainstem, *dmPFC* dorsomedial prefrontal cortex, *IFG* inferior frontal gyrus, *MTG* middle temporal gyrus, *PCC* posterior cingulate cortex, *SMA* supplementary motor area, *TPJ* temporo-parietal junction (cf. Bzdok et al. 2012)

and the right AM in empathy, that is, the same area in contralateral hemispheres. A role of the AM in moral cognition is supported by correlation between its neural activity and the participants' level of self-reported emotional arousal when presented with visual stimuli of people being harmed (Decety et al. 2012). More specifically, it is known that AM activity typically increases in the left hemisphere in controlled, elaborate social-cognitive processes and in the right hemisphere in automatic, basic emotional processes (Markowitsch 1998; Phelps et al. 2001). This lateralization pattern potentially explains the consistent engagement of the left AM in moral cognition (more controlled/elaborate) and right AM in empathy (more automatic/basic). Furthermore, activity in the PCC was found in adjacent, yet nonoverlapping, locations during moral cognition and empathy. Neural activity in this brain area was observed in hearing and recalling affective autobiographical episodes, dealing with coherent social scenarios, viewing familiar faces, as well as emotional planning. The PCC was thus repeatedly proposed to integrate retrieval of past experiences and ongoing emotion processing, which is potentially shared by moral cognition and empathy.

To sum up the meta-analytic evidence, some aspects of affective processing are probably shared by moral cognition and empathy, as the respective meta-analyses

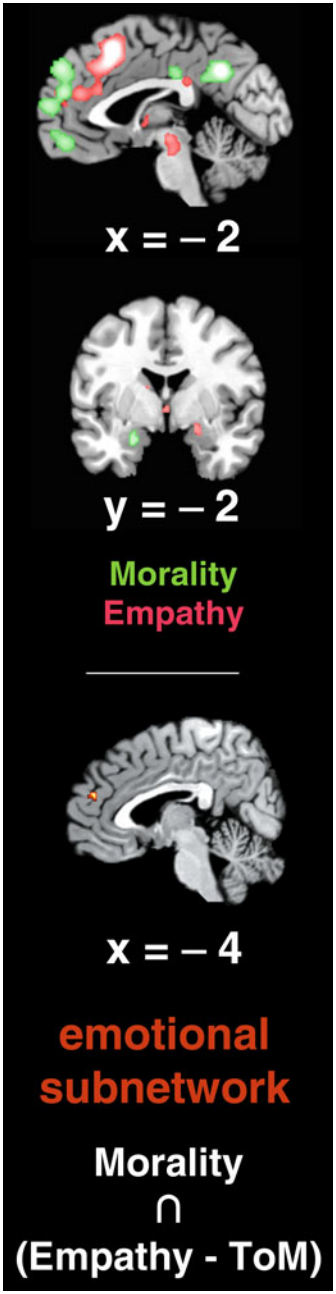

Fig. 9.5 Relationship between the moral network and the neural network underlying empathy. *Bottom panel*: overlapping activation patterns between the meta-analysis on moral cognition and the difference analysis between empathy and ToM. *Top panel*: sagittal and coronal slices of juxtaposed results from the meta-analyses on moral cognition and empathy to highlight similar convergence in the posterior cingulate cortex and amygdala. Coordinates in MNI space (cf. Bzdok et al. 2012)

revealed convergence in the dmPFC (direct overlap), AM (oppositely lateralized), and PCC (closely adjacent clusters). A single direct overlap in the dmPFC might further suggest representational socio-emotional processing to be common to moral cognition and empathy. More generally, it is interesting to note that the neural correlates of moral cognition are more closely related to the neural signature of ToM than to that of empathy. It is an important question whether this is an epiphenomenon of methodological idiosyncracies or illustrates brain network dynamics in everyday life.

The Neurobiological Relationship Between Moral Cognition and Mind-Wandering

It is becoming increasingly clear that brain areas pertaining to social cognition are topographically highly similar to brain areas that increase activity in the idling, unconstrained mind, the so-called "default mode," and decrease activity during stimulus-driven, goal-directed tasks (Shulman et al. 1997; Spreng et al. 2009). More specifically, brain areas underlying unconstrained cognition were consistently associated with a number of complex, introspective mental tasks, including contemplating mind states, self-focused reflection, mental navigation of the body in space, autobiographical memory recall, and, more generally, envisioning situations detached from reality. Performing separate meta-analyses on moral and unconstrained cognition hence allowed elucidating the correspondence between the neural substrates consistently engaged by these two mental states. We included those neuroimaging experiments from the BrainMap database (Laird et al. 2011) into the meta-analysis of unconstrained cognition whose metadata indicated them to provide coordinates of brain deactivation (Fig. 9.6). Brain activity patterns during moral reasoning and unconstrained cognition overlapped in the vmPFC, dmPFC, precuneus, and bilateral TPJ (Fig. 9.7). Consequently, those brain areas consistently implicated in moral cognition indeed lower their activity during stimulus-driven, goal-directed cognition. More broadly, the observed similarities of the neural networks underlying moral and unconstrained cognition favor a possible relationship between the physiological baseline of the human brain and a psychological baseline implicated in constant social cognition.

What is the common denominator of moral and unconstrained cognition? It was speculated that the human brain might have evolved to, by default, predict environmental events using mental imagery. In particular, autobiographical memory supplies building blocks of social semantic knowledge. Isolated conceptual scripts may then be reassembled to enable forecasting future events (Tulving 1983). Constructing detached probabilistic social scenes could thus influence perception and behavior by estimating saliency and action outcomes (Boyer 2008; Schilbach et al. 2008). Ultimately, the tonically active default mode network might be adapted to gathering sensory information for the probabilistic mapping of the external world in order to optimize the organism's behavioral response. That is,

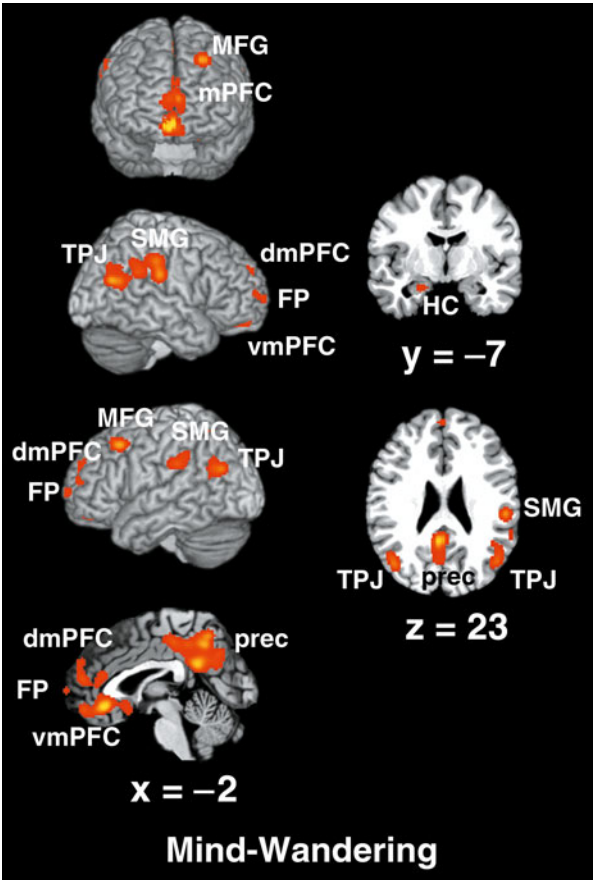

Fig. 9.6 Meta-analysis results on mind-wandering. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to mind-wandering (533 neuroimaging experiments), i.e., brain activity in the absence of a specific task. Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *dmPFC* dorsomedial prefrontal cortex, *HC* hippocampus, *FP* frontal pole, *MFG* middle frontal gyrus, *mPFC* medial prefrontal cortex, *prec* precuneus, *SMG* supramarginal gyrus, *TPJ* temporo-parietal junction, *vmPFC* ventromedial prefrontal cortex (Schilbach et al. 2012)

pondering over morally relevant social situations and simply letting the mind float might both imply contemplation of hypothetical social scenes that guide actual behavior.

Conclusion and Future Directions

The conjunction of the above quantitative and qualitative reviews suggests that moral cognition might rely in large extent on neural systems related to social cognition. The group of social processes subserving moral cognition appears to include contemplating mind states, emotion processing, and internally directed cognitions, such as self-referential thought, mental imagery, and perhaps even prediction generation. Additionally, the neural correlates of moral cognition were dissociated into more rational and more emotional subsystems by reference to

Fig. 9.7 Relationship between the moral network and the neural network underlying mind-wandering. Overlapping activation patterns between the meta-analyses on moral cognition and mind-wandering, i.e., brain activity in the absence of a specific task. Coordinates in MNI space (cf. Bzdok et al. 2012; Schilbach et al. 2012)

a socio-cognitive framework (ToM cognition) and a socio-affective framework (empathy), respectively. The neural network underlying moral cognition might thus be considered “domain-global” as it reflects functional integration of distributed brain networks. Put differently, no part of the brain is uniquely devoted to moral cognition but this capacity is very likely deployed across several heterogeneous functional domains. This contention concurs with the observation that there is also no single brain region specific for the development of antisocial/psychopathic behavior deficient in moral cognition (Raine and Yang 2006).

Shifting from the neural level to psychological concepts, the question “what brain areas support moral judgment?” might already be ill-posed, or at least inaccurate, because moral cognition is unlikely to be a unified psychological entity. That is, distinct (sets of) cognitive components probably support distinct classes of moral judgments. Even more fundamentally, “morality” as such might actually not be localizable in the brain at all given that it constitutes a complex cultural concept, that is, a phenomenon of human cultural evolution. Considering from a phenomenological, psychological, and philosophical point of view, there might even be nothing unique to the notion of “morality” itself. What is unique to moral judgments comparing to, for example, judgments of complex social scenarios? In short, we cannot measure “morality” itself in the human brain. Instead, we can measure brain activity of individuals lying in a neuroimaging scanner while thinking about moral issues. For these reasons, it might actually be naive to search for something like a “distinct moral module” in the first place. Furthermore, moral psychology and moral neuroscience mainly concentrated on moral decision-making, rather than the very manifestation of morality – moral behavior (see Moll et al. 2006 for an exception).

From a methodological perspective, it is unclear to what extent existing neuroimaging studies on moral cognition suffer from this potentially limited ecological validity. That is, the used experimental tasks might only partially involve the neural processes that navigate real-life moral behavior. In particular, complicated dilemmas borrowed from moral philosophy were often employed as stimulus material. This is epitomized by the “trolley dilemma” that prompts a decision between either letting five people die or actively causing the death of another single person to save the life of those five people. A tendency for artificial moral scenarios, on top of the experimental constraints of neuroimaging environments, could have entailed a systematic overestimation of cognitive versus emotional processes. The observed bigger correspondence of the moral neural network with that of ToM, rather than empathy, might thus be epiphenomenal of established experimental features.

Moreover, neuroimaging results were often discussed by qualitative comparison between studies that differ in various crucial aspects, including stimulus material (text vs. pictures/movies vs. numbers in neuroeconomic games), the participants’ perspective (engaged second-person vs. observant third-person perspective), control conditions (morally relevant vs. morally irrelevant, high-level vs. low-level), or continuity (single-shot judgments vs. multi-trial paradigms). It is conceivable that the neural instantiation of moral judgments and comparisons between those are highly susceptible to such experimental differences, especially in light of the integrative character of moral cognition.

From a neuroethical perspective, it can be said that the neuroscience of moral decision-making may be able to contribute to ethics by providing descriptive results. But one must be wary of overextending the logic of neuroimaging with regard to morality. Although moral decision-making correlates with the activation of specific regions in the brain, this does not necessarily mean that moral judgment can be reduced to this activation. A central problem is thus that an attempt is made under scientific conditions to establish which brain areas are particularly active during moral cognition without it being possible to define exactly what moral cognition as such is and without the existence of an objectifiable moral theory (Gazzaniga 2007).

Some critics even raise much more fundamental objections to imaging studies of moral cognition. They pose the rhetorical question: How can I hope at all to discover "facts" which prove "values"? All moralities and all moral decisions are in fact based on values and norms, and these cannot, or at least cannot necessarily be reduced or ascribed for their part to (neurophysiological) facts but are intrinsically subjective and hence may only be accessible on phenomenological accounts. A second question is equally important: What is the consequence if I regard and acknowledge specific activities of neuronal tissue as the basis for moral evaluations and decisions? If values and norms can be described as sequences of cellular processes, does this not remove the basis for morality? (Vogelsang 2008).

Neil Levy (2007) attempted to summarize this "challenge from neuroscience to morality" in four consecutive sentences:

1. Our moral theories, as well as our first-order judgments and principles are all based, more or less directly upon our moral intuitions.
2. These theories, judgments and principles are justified only insofar as our intuitions track genuinely moral features of the world.
3. But our moral intuitions are the product of cognitive mechanisms which evolved under non-moral selection pressures and therefore cannot be taken to track moral features of the world; hence
4. Our moral theories, judgments and principles are unjustified.

It emerges from the above that the study and objectification of processes involving moral decisions and judgments poses fundamental problems. The ability to image brain activity is much greater than the ability to draw clear conclusions with regard to questions of morality from it.

Critics also object that it is not clear to them why knowledge about neuronal processes should help us at all when it comes to morality and moral decisions. They point out that human morality – beyond all modern empirical methods of access – has always been a main topic of philosophical debate and that this debate will continue to be needed in the future (Brukamp and Groß 2010; Groß 2010). Particularly when normative conclusions are to be made on the basis of these empirical findings, the problem of the "ought" fallacy arises, as a (moral) "ought" cannot simply be derived from a (neuronal) "is." This also makes it necessary to raise the question of the scientific and normative rules for dealing with scientific studies on moral cognition. What standards must be met by the scientists who carry

out and publish studies of this kind? Is it enough for them to be experts in neuroscience esp. neuroimaging or should (additional) expertise in the field of morality and (neuro)ethics be required?

Irrespective of the abovementioned questions and concerns, detailing the neurobiological nature of moral cognition is a goal worth pursuing. Importantly however, moral neuroscience should strive for an explanation of the understanding of morality which underlies it, for realistic moral scenarios and for a more rigorous across-study discussion (cf. Knutson et al. 2010). These suggestions might help to minimize the risk of investigating "in vitro moral cognition."

Cross-References

▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
▶ Mental Causation
▶ Moral Cognition: Introduction
▶ Moral Intuition in Philosophy and Psychology
▶ The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy

References

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences, 1191*(1), 42–61. doi:10.1111/j.1749-6632.2010.05445.x. NYAS5445 [pii].

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience, 2*(11), 1032–1037. doi:10.1038/14833.

Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W. L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: Evidence for consistent performance on non-verbal, "reality-unknown" false belief and false photograph tasks. *Cognition, 103*(2), 300–321. doi:10.1016/j.cognition.2006.04.012. S0010-0277(06)00082-5 [pii].

Boyer, P. (2008). Evolutionary economics of mental time travel? *Trends in Cognitive Science, 12*(6), 219–224. doi:10.1016/j.tics.2008.03.003. S1364-6613(08)00093-4 [pii].

Brukamp, K., & Groß, D. (2010). Neuroenhancement – A controversial topic in contemporary medical ethics. In P. A. Clark (Ed.), *Contemporary issues in bioethics* (pp. 39–50). [Ebook]. www.intechopen.com/books/contemporary-issues-in-bioethics/neuroenhancement-a-controversial-topic-in-medical-ethics. Rijeka

Bzdok, D., Langner, R., Hoffstaedter, F., Turetsky, B. I., Zilles, K., & Eickhoff, S. B. (2012a). The modular neuroarchitecture of social judgments on faces. *Cerebral Cortex, 22*(4), 951–961. doi:10.1093/cercor/bhr166. bhr166 [pii].

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012b). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function, 217*(4), 783–796. doi:10.1007/s00429-012-0380-y.

Bzdok, D., Laird, A., Zilles, K., Fox, P. T., & Eickhoff, S. (2012, in press). An investigation of the structural, connectional and functional sub-specialization in the human amygdala. *Human Brain Mapping*. doi:10.1002/hbm.22138

Bzdok, D., Langner, R., Schilbach, L., Engemann, D. A., Laird, A. R., Fox, P. T., Eickhoff, S. B. (2013). Segregation of the human medial prefrontal cortex in social cognition. *Front Hum Neurosci, 29*(7), 232. doi: 10.3389/fnhum.2013.00232.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129*(Pt 3), 564–583. doi:10.1093/brain/awl004. awl004 [pii].

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*(2), 71–100. doi:10.1177/1534582304267187. 3/2/71 [pii].

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex, 22*(1), 209–220. doi:10.1093/cercor/bhr111. bhr111 [pii].

Eickhoff, S. B., & Bzdok, D. (2012). Meta-analyses in basic and clinical neuroscience: State of the art and perspective. In S. Ulmer & O. Jansen (Eds.), *fMRI – Basics and clinical applications* (2nd ed.). Heidelberg: Springer.

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping, 30*(9), 2907–2926. doi:10.1002/hbm.20718.

Eisenberger, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology, 51*, 665–697.

Ferstl, E. C., & von Cramon, D. Y. (2002). What does the frontomedian cortex contribute to language processing: Coherence or theory of mind? *NeuroImage, 11*, 1599–1612.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358*(1431), 459–473. doi:10.1098/rstb.2002.1218.

Gazzaniga, M. S. (2007). *Wann ist der Mensch ein Mensch? Antworten der Neurowissenschaft auf ethische*. Fragen: Patmos.

Gitelman, D. R., Nobre, A. C., Parrish, T. B., LaBar, K. S., Kim, Y. H., Meyer, J. R., & Mesulam, M. (1999). A large-scale distributed network for covert spatial attention: Further anatomical delineation based on stringent behavioural and cognitive controls. *Brain, 122*(Pt 6), 1093–1106.

Gorno-Tempini, M. L., Rankin, K. P., Woolley, J. D., Rosen, H. J., Phengrasamy, L., & Miller, B. L. (2004). Cognitive and behavioral profile in a case of right anterior temporal lobe neurodegeneration. *Cortex, 40*(4–5), 631–644.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science, 6*(12), 517–523. S1364661302020119 [pii].

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. doi:10.1126/science.1062872. 293/5537/2105 [pii].

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400. doi:10.1016/j.neuron.2004.09.027. S0896627304006348 [pii].

Groß, D. (2010). Traditional vs. Modern neuroenhancement. Notes from a medico-ethical and societal perspective. In H. Fangerau & T. Trapp (Eds.), *Implanted mind (= Science Studies)* (pp. 137–157). Transcript: Bielefeld.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. doi:10.1037//0033-295x.108.4.814.

Hare, R. D. (2003). *The Hare psychopathy checklist – Revised* (2nd ed.). Toronto: Multi-Health Systems.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus – It's my area: Or is it? *Journal of Cognitive Neuroscience, 20*(12), 2125–2136. doi:10.1162/jocn.2008.20148.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*(12), 2949–2957. doi:10.1016/j.neuropsychologia.2008.06.010. S0028-3932(08)00261-3 [pii].

Knutson, K. M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J. R., Vasudeva, V., & Grafman, J. (2010). Behavioral norms for condensed moral vignettes. *Social Cognitive and Affective Neuroscience, 5*(4), 378–384. doi:10.1093/scan/nsq005. nsq005 [pii].

Laird, A. R., Eickhoff, S. B., Fox, P. M., Uecker, A. M., Ray, K. L., Saenz, J. J., Jr., & Fox, P. T. (2011). The brainmap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Research Notes, 4*(1), 349. doi:10.1186/1756-0500-4-349. 1756-0500-4-349 [pii].

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184. doi:10.1146/annurev.neuro.23.1.155.

Levy, N. (2007). *Neuroethics. Challenges for the 21st century*. Cambridge: Cambridge University Press.

Maddock, R. J. (1999). The retrosplenial cortex and emotion: New insights from functional neuroimaging of the human brain. *Trends in Neurosciences, 22*(7), 310–316. S0166223698013745 [pii].

Margulies, D. S., Vincent, J. L., Kelly, C., Lohmann, G., Uddin, L. Q., Biswal, B. B., & Petrides, M. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences of the United States of America, 106*(47), 20069–20074. doi:10.1073/pnas.0905314106. 0905314106 [pii].

Markowitsch, H. J. (1998). Differential contribution of right and left amygdala to affective information processing. *Behavioural Neurology, 11*(4), 233–244.

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex, 18*(2), 262–271. doi:10.1093/cercor/bhm051. bhm051 [pii].

Moll, J., & Schulkin, J. (2009). Social attachment and aversion in human moral cognition. *Neuroscience and Biobehavioral Reviews, 33*(3), 456–465. doi:10.1016/j.neubiorev.2008.12.001. S0149-7634(08)00204-2 [pii].

Moll, J., de Oliveira-Souza, R., & Eslinger, P. J. (2003). Morals and the human brain: A working model. *Neuroreport, 14*(3), 299–305. doi:10.1097/01.wnr.0000057866.05120.28.

Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignacio, F. A., Bramati, I. E., Caparelli-Daquer, E. M., & Eslinger, P. J. (2005a). The moral affiliations of disgust: A functional MRI study. *Cognitive and Behavioral Neurology, 18*(1), 68–78. 00146965-200503000-00008 [pii].

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005b). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*(10), 799–809. doi:10.1038/nrn1768. nrn1768 [pii].

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America, 103*(42), 15623–15628. doi:10.1073/pnas.0604475103. 0604475103 [pii].

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 2688–2692. doi:10.1073/pnas.1011734108. 1011734108 [pii].

Olson, I. R., Ploaker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain, 130*, 1718–1731. doi:10.1093/Brain/Awm052.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience, 4*(4), 437–441. doi:10.1038/86110. 86110 [pii].

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *The Behavioral and Brain Sciences, 1*(4), 515–526.

Raine, A., & Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. *Social Cognitive and Affective Neuroscience, 1*(3), 203–213. doi:10.1093/scan/nsl033.

Ross, L. A., & Olson, I. R. (2010). Social cognition and the anterior temporal lobes. *NeuroImage, 49*(4), 3452–3462. doi:10.1016/j.neuroimage.2009.11.012. S1053-8119(09)01196-3 [pii].

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14*(4), 303–316.

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition, 17*(2), 457–467. doi:10.1016/j.concog.2008.03.013. S1053-8100(08)00037-8 [pii].

Schilbach, L., Bzdok, D., Timmermans, B., Fox, P. T., Laird, A. R., Vogeley, K., & Eickhoff, S. B. (2012). Introspective minds: Using ALE meta-analyses to study commonalities in the neural correlates of emotional processing, social & unconstrained cognition. *PloS One, 7*(2), e30920. doi:10.1371/journal.pone.0030920. PONE-D-11-20368 [pii].

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain, 132*, 617–627. doi:10.1093/Brain/Awn279.

Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., & Petersen, S. E. (1997). Common blood flow changes across visual tasks. 2. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience, 9*(5), 648–663.

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences, 1156*, 81–96. doi:10.1111/j.1749-6632.2009.04418.x. NYAS04418 [pii].

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489–510. doi:10.1162/jocn.2008.21029.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345–372.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon.

Vogelsang, F. (2008). Aufgaben und Perspektiven einer künftigen Neuroethik. In F. Vogelsang & C. Hoppe (Eds.), *Ohne Hirn ist alles nichts. Impulse für eine Neuroethik* (pp. 11–22). Neukirchener.

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex, 21*(11), 2461–2470. doi:10.1093/cercor/bhr016. bhr016 [pii].

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*(4), 1912–1920. doi:10.1016/j.neuroimage.2008.01.057. S1053-8119(08)00087-6 [pii].

Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396–1405. doi:10.1162/jocn.2009.21137.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America, 104*(20), 8235–8240. doi:10.1073/pnas.0701408104.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America, 107*(15), 6753–6758. doi:10.1073/pnas.0914826107. 0914826107 [pii].

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 104*(15), 6430–6435. doi:10.1073/pnas.0607061104. 0607061104 [pii].

Psychology and the Aims of Normative Ethics

10

Regina A. Rini

Contents

Abstract

This chapter discusses the philosophical relevance of empirical research on moral cognition. It distinguishes three central aims of normative ethical theory: understanding the nature of moral agency, identifying morally right actions, and determining the justification of moral beliefs. For each of these aims, the chapter considers and rejects arguments against employing cognitive scientific research in normative inquiry. It concludes by suggesting that, whichever of the central aims one begins from, normative ethics is improved by engaging with the science of moral cognition.

R.A. Rini
University of Oxford, Oxford, UK
e-mail: regina.rini@philosophy.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_162,

Three Central Questions of Normative Ethics

It is undeniable that the field of empirical moral psychology has grown dramatically in the last decade, with new experimental techniques allowing us unprecedented understanding of the causal and computational structures of the human moral faculty. Or, at least, it is undeniable that this research contributes to a *descriptive* project, one of better understanding the facts about who we are and how we think (Doris and Stich 2007; Appiah 2008; Levy 2009; Knobe 2010). But what might be denied is that these investigations have much to offer to normative ethics, a distinctively *prescriptive* sort of inquiry.[1] The purpose of this chapter is to show why normative ethics – the study of how we ought to live our lives and what we ought to value – can indeed benefit from engagement with empirical moral psychology.[2]

It makes sense to begin with some conception of what normative ethics *is* and how the inquiry operates. Unfortunately, there is nothing like consensus on those matters among ethicists. Therefore, this chapter pursues a piecemeal dialectical strategy, setting out different ways one might characterize the discipline, and then asking of each in turn how empirical moral psychology might be brought to bear.

Each of the following is a *central question* of normative ethics:

1. What is it to be a moral agent?
2. Which actions are morally permitted or required?
3. Which of our moral beliefs are justified?

These questions are not necessarily rivals; one could certainly hold that normative ethics aims at addressing *all* of these questions. Some philosophers do see one of these questions as conceptually prior to the others, or indeed see one as exhaustively constituting the subject matter of normative ethics, but those disputes will not be engaged here.

Each of the following sections discusses one of these questions, and each section follows a common structure. First will be a *negative* argument, one claiming that answering the target central question allows little relevance for empirical moral psychology. This followed by a reply, arguing that further reflection on the central

[1]There is one sense in which no one doubts the relevance of empirical findings to normative ethics. That is in *applying* settled normative views to actual circumstances. Obviously psychology – and ordinary daily experience – can aid us in learning how to bring about the moral goals we have set, once those goals are *already* determined. What is at issue in this chapter is something different: Can empirical moral psychology play a role in helping to determine what the moral goals themselves ought to be?

[2]Some clarification about terms: The use of “empirical moral psychology” is meant to be ecumenical, encompassing research by psychologists, neuroscientists, biologists, behavioral economists, sociologists, and experimental philosophers. “Normative ethics” refers to the branch of moral philosophy concerned with how we ought to live our lives, what things we ought to value, and what practical decisions we ought to make. This chapter does *not* discuss certain related topics, such as free will and moral responsibility, or naturalistic moral ontology.

question instead favors a more welcoming conclusion for empirical moral psychology. Finally, there are illustrative examples of relevant empirical research. The point of this structure is partly expository, since the chapter aims to provide an overview of many branches of a growing debate. But there is also a dialectical strategy here: If it can be shown that, *whichever* of the three central questions one begins from, negative arguments can be reversed into positive arguments, then the relevance of empirical moral psychology should be convincingly secured.

What Is It To Be a Moral Agent?

Challenge: Rational Agency and the Practical Perspective

Most healthy adult human beings are moral agents, capable of engaging with moral reasons and being held morally responsible for their actions. But what constitutes a moral agent? Many philosophers have thought that answering this question is central to normative ethics, in that providing an answer would subsequently guide theory about how to live our lives. This section begins by sketching one very influential answer – one which appears to have the consequence of rendering empirical psychology irrelevant to normative ethics.

The Enlightenment philosopher Immanuel Kant claimed that basic elements of normative ethics (such as his famous categorical imperative) follow directly from a conception of moral agency as grounded in *rational nature*.[3] According to Kant, taking moral decisions seriously requires thinking of ourselves as rational beings, as beings who can determine for ourselves the rules that govern our actions, rather than being determined by the laws of nature that govern everything else. When we take up the question of how we should live our lives, it makes no sense to think of ourselves in terms of natural laws. Asking what we *should* do presupposes that we have some say in the matter, whereas an investigation of natural laws aims at showing what *must* happen. So for Kant, moral inquiry is fundamentally *practical*, in that it is conducted from an orientation aimed at guiding genuine choices, and not simply at describing the world.

Clearly, on this conception of moral agency, there will be difficulties employing empirical moral psychology in normative ethics. Psychology concerns itself with the causal structure of our thought; in Kant's terms, it aims at describing the natural laws that govern how we think. But describing causal structures *cannot be* the same project as deciding how we ought to live our lives, which takes place from the practical perspective. So Kant was highly critical of theorists who attempted to

[3]The clearest statement of Kant's view on this point comes in Book 3 of his *Groundwork for the Metaphysics of Morals* (Kant 1785). Interpreting Kant is always a delicate matter, and there is not space here to fully explicate the interpretation given in the text, which mostly follows Korsgaard (1996a).

draw moral conclusions from contingent facts about the human mind and its circumstances. Instead, he thought, we must pursue "a pure moral philosophy which is fully cleansed of everything that might be in any way empirical and belong to anthropology."[4] Many contemporary philosophers follow Kant here, arguing that psychological approaches simply confuse the practical point of moral inquiry. So, writes Thomas Nagel, "The pursuit of objective practical principles is not to be conceived of as a psychological exploration of our moral sense, but as an employment of it (Nagel 1986, pp. 150–151)". (See also Nagel 1978 and Fried 1978.)

Reply: Reflective Agency and Self-Understanding

It seems best to concede the core of Kant's criticism: There is something distinctive about conceiving of ourselves as moral agents, and this does not sit well alongside a psychological self-conception. But, unless we adopt a very radical sort of metaphysics, it seems we must still admit that, in the end, we *are* subject to the laws of nature, including psychological laws. Kant need not deny this: His point is simply that there is a problem in *simultaneously* thinking in this way and thinking of ourselves from the practical perspective.[5] The two perspectives, practical and psychological, cannot be entered into at the same time.

This is not a particularly satisfying conclusion. It makes the practical perspective look a bit like a petulant child: covering her eyes, plugging her ears, and refusing to acknowledge the presence of the laws of nature all around her. And it makes the psychological perspective sound myopic and sterile, divorced from the things we actually care about in life. The problem here comes from Kant's refusal to engage with contingent, empirical facts about human nature: He insists upon seeing a moral agent strictly as a *rational* agent, with no other attributes. Could we dispense with this aspect of Kant's approach, permitting consideration of some contingent elements of our nature, without abandoning the essential point of the practical perspective?

A very appealing approach emphasizes that we are *reflective* entities. Unlike mechanical devices or simple animals, we can think about the motives of our thoughts and actions, and it is essential to our nature that we have the ability to reflectively endorse or refrain from endorsing these motives (Frankfurt 1971). Christine Korsgaard, in developing her contemporary Kantianism, has taken

[4]Kant (1785), *Ak* 4:389. Kant is quite strident on this point; he goes on to insist that an empirical approach to fundamental moral principles gives only a "base way of thinking," "disadvantageous to the purity of moral themselves… a bastard patched together from limbs of quite diverse ancestry" (4:425–426).

[5]Kant often certainly sounds as if he *is* making a metaphysical claim, where adopting the practical perspective entails denying that the world really contains deterministic natural laws. But this is not the only interpretation available. See Korsgaard (1996a).

a leading role in highlighting the role of reflection in moral agency. Korsgaard offers what she calls the *transparency* requirement on normative ethics:

> A normative moral theory must be one that allows us to act in the full light of knowledge of what morality is and why we are susceptible to its influences, and at the same time to believe that our actions are justified and make sense.[6]

The reflective perspective, unlike Kant's practical perspective, is not incompatible with acknowledging contingent facts about our nature. It need only insist that our focus must be on *evaluating* these natural facts, rather than merely describing them. Understood this way, the reflective perspective does not at all reject the relevance of empirical psychology to normative ethics. In fact, it supports an argument *requiring* psychological inquiry:

The Self-Understanding Argument

1. Being moral *agents* requires that we understand and endorse the motives of our judgments and actions.
2. Many of the motives of our judgments and actions can only be fully understand in light of findings from empirical moral psychology.
3. Therefore, in order to be effective moral agents, we must pay attention to discoveries of empirical moral psychology.

I take it that step (1) is granted by anyone who prefers the gentler reflective perspective over Kant's mysterious practical perspective.[7] Step (2) is best shown through example.

Examples: Human Nature and Normative Ethics

If one wants evidence for how empirical psychology might aid in reflective self-understanding, it could help to look away from Kant for awhile. Another deep tradition in moral theory sees morality as a natural phenomenon, growing out of basic facts about the biological constitution of human beings and their relations to one another (Foot 2003; Kitcher 2011). Owen Flanagan (1996) traces this tradition from Aristotle through John Dewey, and argues that informing moral philosophy through contemporary psychology is another, necessary iteration. The traditional Aristotelian approach assumed that there is an ultimate *telos*, or purpose, to human nature, that the essence of being human is aiming at a state

[6](Korsgaard 1996b, p. 17). For a related discussion specifically regarding psychological findings, see Kamm (2009, p. 469).

[7]Although objections to step (1) are possible, one might challenge the idea of agency itself, or at least the idea that reflective endorsement is a necessity for it. Doris (2009) makes an argument of this sort. Alternately, one might suggest that the connection between steps (1) and (2) is less clear. See van Roojen (1999) for an argument that the kinds of reasons relevant to moral agency do not map onto the kinds of motives discussed in psychological theory.

of perfected humanness. One need not take teleology on board to adopt this sort of view. The key point is simply that understanding how we ought to live our lives begins with understanding what sorts of creatures we are, in our capacity as rational agents *and* as contingent, limited organisms. When we attend to empirical discoveries, we may find grounds for reevaluating tenets of normative theory.

Start, for instance, with biology. The primatologist Frans de Waal argues that social contract theory (the sort advocated by Hobbes and Rawls) assumes that the construction of complex social arrangements is needed to mitigate the effects of individual self-interest. However, de Waal claims, careful study of primate behavior and the evolutionary origin of the human species would instead lead to the conclusion that we are predisposed to social cooperation; we are "obligatorily gregarious" (de Waal 2006, p. 4). If this is recognized, then one of the central puzzles of moral inquiry becomes inverted: Instead of trying to explain why we ever cooperate, we should instead focus on understanding why we sometimes fail to do so. This explanatory inversion does not eliminate the need to reflectively evaluate our altruistic practices, but it does suggest that there is far less distance between our natures and our norms than many had assumed (see also Petrinovich et al. 1993 and Sripada 2005).

Now consider social interaction. Gossip is often thought of as morally extraneous: idle chatter aimed purely at titillation, rather than at seriously evaluating the ethicality of particular actions. But according to (Sabini and Silver 1982, p. 101), who argue partly on evidence from social psychology, gossip in fact fulfills an essential role in moral practice. Gossip, they say, functions to coordinate social norms, and does so in a way permitting far greater detail and nuance than what is possible in formal moral instruction or debate. If this is right, we might wish to reevaluate how we regard gossipers. And, more importantly, normative ethics might benefit from examining moral commitments disclosed through informal gossip, alongside more traditional moral intuitions.[8]

Two important and ambitious research programs provide wide-ranging treatments of our psychological moral nature. One is the developmental tradition pioneered by Jean Piaget, and extended by Lawrence Kohlberg.[9] According to this account, human moral judgment arises from an invariant sequence of developmental stages, each consisting of logical improvements upon the limitations of prior stages. Kohlberg regards this account as capturing the essence of the moral domain so completely that, he says, "an adequate psychological analysis of the structure of

[8]Similar comments apply to the phenomenon of moral *disagreement*. Knowing how and why people come to hold diverging moral views – between various communities (Moody-Adams 2002; Haidt 2012) or even within individual minds (Cushman and Young 2009) – might provide clues as to how to deal with them.

[9]Piaget, in fact, aims his work squarely at addressing Kant's challenge to empirical psychology. Piaget's developmental account is explicitly intended to reveal the nature of moral agency (or autonomy). For a development of Piaget's views on normative ethics, see Rini (unpublished manuscript).

a moral judgment, and an adequate normative analysis of the judgment will be made in similar terms."[10]

A second, more recent, research program concerns the interpretation of moral intuitions. Reflective equilibrium, the dominant method of contemporary normative ethics, involves soliciting intuitive reactions to test cases, which are systematized in search of latent principles. The method is most fully articulated in the work of John Rawls (1951, 1971), where it is primarily presented as a means of *justifying* moral principles. But Rawls himself noted a resemblance to the *descriptive* enterprise of linguistic psychology (Rawls 1971, p. 47). A number of authors (Mikhail 2011; Dwyer 2006; Hauser 2006) have pressed this *linguistic analogy*, arguing that moral intuitions result from a domain-specific, partly innate cognitive, mechanism, by which our minds map the causal structure of observed situations onto morally valenced reactions. The suggestion is that if reflective equilibrium has an aim of finding the underlying structure of our moral intuitions, then surely the powerful empirical techniques of psychology can be an aid to purely introspective theorizing.[11]

All of these examples are susceptible to empirical challenge, and certainly not everyone is persuaded by each. But the overall point should be clear: To the extent that reflective self-understanding plays a central role in normative ethics, empirical psychology enhances rather than detracts from this role. The claim is not that descriptive psychological findings lead immediately to substantive normative conclusions. Rather, the claim is that a reflective approach to normative ethics unnecessarily hobbles itself if it refuses to engage with this powerful form of self-understanding.

Which Actions Are Morally Permitted or Required?

Challenge: The Is-Ought Gap

In his *Treatise of Human Nature*, David Hume famously expressed surprise at finding other authors moving from claims about what *is* the case to claims about what *ought to be* the case. According to Hume, it "seems altogether inconceivable, how this new relation ['ought'] can be a deduction from others, which are entirely

[10](Kohlberg 1971, p. 224). For a critical appraisal of Kohlberg's normative claims, see Goodpaster (1982). Kohlberg's findings have also been criticized on empirical grounds, especially by his former collaborators (Gilligan 1982) and (Turiel 1983). For more recent work in the developmental tradition, see Rest et al. (1999) and Narvaez and Lapsley (2009).

[11]The linguistic analogy has been criticized both for its normative claims (Daniels 1980) and its empirical grounding (Dupoux and Jacob 2007; Prinz 2008). For general discussions of descriptive interpretations of reflective equilibrium, see Scanlon (2002) opposed, and Rini (2011, Chap. 3), in favor. A closely related research program, focused on describing the causal role of intention-ascription in moral judgment, provides a detailed example of what careful empirical work can uncover (Knobe 2003; Young and Saxe 2008; Cushman 2008).

different from it. (Hume 1739, Book III, sec 1, part 1)" Hume here articulated the *is-ought gap*: the doctrine that descriptive claims (about how matters actually are) constitute a logically different sort than prescriptive claims (about how matters should be), and therefore that something goes wrong when one attempts to generate prescriptive claims from purely descriptive ones.

The is-ought gap is related to a central aim of moral philosophy: identifying which *actions* ought to be pursued. For Hume, the gap stemmed from the idea that moral claims are intrinsically motivational in a way that descriptive claims are not. During the twentieth century, the is-ought gap was frequently formulated as a point about moral *language* (Moore 1903; Mackie 1977, pp. 64–73; Joyce 2006, pp. 146–156). According to a particularly influential mid-century view, moral vocabulary has an imperative logic, even when its surface grammar appears descriptive. For instance: "you have an obligation to help the needy" appears to *describe a fact* about you (your having a thing, an obligation), but it is actually logically similar to the command, "you, go help the needy." If this view is correct, then it is a misuse of language to suggest that moral claims follow from factual claims in any logical sense.[12] Imperatives do not logically follow from declaratives.

Whatever its grounding, the is-ought gap has an intuitive plausibility. The question of what we should do, of what sorts of actions we should engage in, is simply not answered by a purely descriptive characterization of our existing tendencies to engage in or avoid particular actions – and these are precisely the sorts of things studied by empirical moral psychology. There is, then, at least a prima facie burden to explain how psychological inquiry can produce any conclusions relevant to "ought" considerations.

Reply: Ought Implies Can

Alongside the is-ought gap sits another venerable philosophical dictum: "ought implies can." It does not make sense to claim that one "ought" to do something unless it is possible for the person to do so (though see Graham 2011 for a recent opposed view). For instance: It is nonsense to insist that you, personally, ought to stop a dangerous meteor from striking the Earth, because individual people simply do not have this ability. If moral theory aims at providing guidance to actual ethical decisions, then it must respect the limitations of actual human ability.

Applications of "ought" implies "can" are fairly obvious in cases of temporal or physical impossibility. The present argument aims to draw attention to less obvious cases of *psychological* impossibility. Our minds are limited in various ways, and a genuinely action-guiding normative ethics should be attentive to

[12]For examples of the linguistic formulation of the is-ought gap, see Stevenson (1944, pp. 271–276) and Hare (1952, pp. 17–55). Williams (1985, pp. 121–131) argues that the linguistic formulation does not clarify matters.

these limitations. This is the kernel of Owen Flanagan's *Principle of Minimal Psychological Realism*:

> Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for creatures like us.[13]

If we accept this Principle, then clearly empirical moral psychology will be essential to the aims of moral philosophy. Our best clues about our psychological limitations will surely come from empirical inquiry. Moral theory constructed in ignorance of these limitations will issue frustratingly unfulfillable prescriptions.

It is worth distinguishing two ways in which we could be psychologically incapable of meeting some particular moral prescription. The first way is *cognitive*: The prescription requires us to make decisions that involve resources of calculation, memory, or imagination which we do not possess. Consequentialist moral theorists have long been challenged to explain how it could be possible for actual agents to recognize, let alone calculate, the mushrooming consequences of their individual actions, to which they have developed complex responses (Sidgwick 1907; Hooker 1990). But there may be more subtle cognitive limitations on our moral thinking. For instance, Horgan and Timmons (2009) argue that the "frame problem" from cognitive science – concerning the difficulty of providing a discrete specification for the domain of any particular decision – shows that our moral psychology cannot consist solely in a set of general, exceptionless rules, and therefore any moral theory that does consist in such rules violates the "ought"/"can" stricture.[14]

The second way we may be psychologically limited concerns our *motivations*. It may be that we have strong, persistent motivations to behave or refrain from behaving in particular ways. A moral theory that requires us to do things we cannot feel motivated to do is not much use for action-guidance. Petrinovich et al. (1993) and Petrinovich and O'Neill (1996) argue that evolutionary psychology can uncover "biological universals" among human motivations, which are likely to be strict limitations, given their hard-wiring in our adaptive history. Psychological limitations can also play a role in evaluation of social practices. So, writes John Rawls: "It is... a consideration against a conception of justice that in view of the laws of moral psychology, men would not acquire a desire to act upon it even when the institutions of their society satisfied it. For in this case there would be difficulty in securing the stability of social cooperation."[15]

---

[13]Flanagan 1993, p. 32. See also a related argument in Appiah (2008, pp. 22–23).

[14]See also Gigerenzer (2008, p. 6) and Moody-Adams (2002, p. 140) for other applications of "ought"/"can" to limitations on cognitive or imaginative human capacities.

[15](Rawls 1971, p. 138). Rawls continued to draw attention to this role for psychological findings in his later work (Rawls 1974, p. 294, Rawls 1987, p. 24). Rawls' attention to psychological realism has brought his method criticism as overly conservative, but an ought-implies-can argument like the one sketched in this section may actually show this sort of conservatism to be a theoretical virtue. See Rini (2011, Chap. 3).

A strong objection can be raised against the argument from psychological possibility, especially in its motivational form: Often, when we claim that some act is not psychologically possible for us, we are merely using psychology to invent *excuses* for not honoring difficult or unpleasant moral duties. Without question, this is a concern that we ought to address for any particular application of this argument. But then it really will come down to the empirical details, to determine just how malleable our cognitive or motivational capacities are (Schleim and Schirmann 2011). So this point is best understood as a cautionary note rather than a direct objection to the argument.[16]

Examples: Virtue and Psychological Limitations

Interestingly, the argument from psychological possibility is illustrated on *both* opposing sides of an ongoing debate over the relationship between cognitive science and virtue ethics. Virtue ethics is a substantive normative theory, due in large measure to Aristotle, which understands morally correct actions and motivations as those possessed by a person with good moral character. Virtues are stable dispositions of character, inclining one to engage in appropriate actions across a range of circumstances. Virtue ethics has experienced a recent revival, especially in the views of those who see it as providing a richer conception of moral life than offered by other theories (Anscombe 1958; MacIntyre 1981; Hursthouse 2002).

One research program claims that central findings from cognitive science "are more consilient with the assumptions of virtue theory than with other theories (Casebeer and Churchland 2003, p. 173)" (See also Clark 1996 and Churchland 2011). While this position is partly supported by positive arguments – from affective neuroscience and the predictions of connectionist neural modeling – one of its central premises is an ought-implies-can argument against rival theories. According to Paul Churchland, consequentialist and deontological normative theories "surreptitiously presuppose a background theory about the nature of cognition, a theory that we now have overwhelming reason to believe is empirically false."[17] The idea is that cognition in general, and so moral cognition in particular, cannot be partitioned into discrete belief-states with unique logical conditions. Since living according to other normative theories *would* require that sort of psychology (at least according to proponents of the argument), those theories are disqualified on ought-implies-can grounds. Virtue ethics, by contrast, is held to be the only normative theory compatible with the purported facts about cognition.

[16]Even Peter Singer, usually quite uncompromising about the demandingness of morality, allows moral theory to bend for certain psychological limitations. For instance, Singer notes that partiality toward family runs contrary to impersonal consequentialist theory. However, he says, familial partiality is so biologically entrenched that it is better to harness it – and so secure reliable concern for local welfare – than to attempt to fight it. See Singer (1981, pp. 33–36).

[17](Churchland 2000, p. 294). This negative argument is related, but not identical, to Horgan and Timmons' (2009) frame problem, discussed above. For further related arguments, see Stich (1993) and Johnson (1996).

A quite different research program suggests the opposite: Evidence from cognitive science *undermines* the tenability of virtue ethics as a normative theory. Robust findings in social psychology, drawn together under the heading of Situationism, suggest that most human behavior is explained not by persistent character traits, but by automatic responses to narrow situational factors (Darley and Batson 1973; Milgram 1973; Haney et al. 1973; Ross and Nisbett 1991). If this is so, then people simply do not have general character traits – such as courage or fidelity – in the way virtue ethics requires (Doris 2002; Harman 1999; for criticism see Kamtekar 2004; Annas 2005; Sabini and Silver 2005). If it is not possible for people to possess the relevant sort of character traits, then a normative theory evaluating them on that basis cannot be of much use.

Debate over the status of virtue ethics continues, with contributions from psychologists and philosophers alike. However, these arguments ultimately resolve, it is instructive to note the implied agreement among their participants, that moral theory must be responsive to the limitations on human agency revealed by psychological inquiry.

Which of Our Moral Beliefs Are Justified?

Challenge: Justification and Explanation

A final central aim of normative ethics is epistemic: We aim to determine which of our moral beliefs (about actions or states of affairs) are *justified*. We presume that some of our existing moral beliefs are *not* justified; we aim to sort out which of these are, and to eliminate them from our normative theories. One way we do this, in reflective equilibrium, involves reflective consideration of moral intuitions solicited by relevant test cases. Those beliefs which best unify intuitions surviving reflective scrutiny are the ones we take to be justified. The others get discarded, and a complete moral theory is constructed only from justified beliefs.

In contrast, a project of simply *describing* our present beliefs cannot be responsive to the question, as it would not discriminate between justified and unjustified beliefs. So it is unclear how empirical moral psychology could have any relevance here, since it is just such a descriptive enterprise. More particularly, empirical psychology aims at *explaining* the causal structure of our beliefs. But, as many philosophers have noted, explaining a belief and justifying a belief are logically separate pursuits. From a first-person perspective, that of the person who "owns" the moral belief, an explanation of its causal structure has no obvious bearing on the matter of whether one ought to go one believing it. Writing in this vein, Ronald Dworkin asks us to imagine that new discoveries show some of our beliefs about justice to be secretly motivated by concerns of self-interest:

> It will be said that it is unreasonable for you still to think that justice requires anything, one way or the other. But why is that unreasonable? Your opinion is one about justice, not about

> your own psychological processes... You lack a normative connection between the bleak psychology and any conclusion about justice, or any other conclusion about how you should vote or act. (Dworkin 1996, pp. 124–125)

As Dworkin's example suggests, psychological findings cannot themselves tell us which of our moral beliefs are justified, because such findings are not *about* the content of moral beliefs. An explanatory inquiry has a different subject matter than a justificatory inquiry (see Strawson 1974, p. 25; Nagel 1997, p. 105).

Reply: The Conditions of Moral Judgment

Immediately following the quotation above, Dworkin concedes that the "bleak psychology" might lead you to rethink your moral beliefs, *if* you added a normative premise stepping from the factual claim to a normative conclusion. For instance, you might decide that aims pursued from covert self-interest are morally unacceptable. Indeed, once such normative premises are included, explanations for beliefs are readily conceded to undermine apparent justifications; philosophers have a notion of "explaining away" intuitively grounded beliefs for precisely this purpose.

In fact, established methodology in normative ethics allows that moral intuitions can sometimes be disqualified on grounds of their *originating conditions*. We sometimes reject an intuition not because of its content (i.e., that some action is right or wrong), but because the thinking that led to the intuition occurred in some concrete situation which we regard as unlikely to generate credible intuitions. So, for instance, we do not trust intuitions originating when we are distracted, tired, intoxicated, etc. Self-interest, indeed, may be another such condition: We think we do well not to trust moral intuitions formed while we are primed to worry about our own stake in some matter.[18]

Notice that the practice of excluding intuitions on grounds of originating conditions does not challenge Dworkin's claim. In these cases, we must still have a normative premise – perhaps implicit – that intuitions formed under conditions of intoxication or self-interest are *unreliable* intuitions, intuitions that cannot convey justification to any belief or theory constructed from them.[19] Excluding intuitions requires stepping beyond an explanatory project and into a justificatory one.

It may be best for proponents of empirical moral psychology to concede all of this, because it shows their program to be simply an extension of existing philosophical methodology. Since it is already accepted that originating conditions are sometimes grounds for disqualifying intuitions, there should be no principled

[18]Rawls stresses that his method of reflective equilibrium is meant to operate only upon "considered moral judgments," which are a class of intuitions rendered according to constraints like these. See Rawls (1951, p. 179) and Rawls (1971, pp. 47–48).

[19]There is a difficult issue here about what *type* of norm plays this role: Intoxicated intuitions might be excluded on epistemic grounds (intoxication being though not conducive to truth in *any* domain), while the self-interest exclusion may represent a distinctively *moral* norm. But such issues can be set aside here.

objection to similar arguments employing psychological premises – provided these arguments also include satisfactory normative premises.

What psychology uniquely offers is an enhanced understanding of the nature of originating conditions. Sometimes it is easy to notice that certain particular intuitions have occurred under conditions of intoxication or self-interest; we can know this through direct self-observation. But other times, facts about the conditions in which our intuitions originate are less obvious to us, such as when we do not realize we are primed for self-interest, or even when we do not recognize the intoxicating influence of some drug or environmental cue. The observational and correlational powers of empirical psychology are likely to be far better at tracking such originating conditions. More importantly, psychology may allow us to come to recognize *new* categories of disqualifying originating conditions, by exposing the causal relations embedded within them.

Examples: Psychological Originating Conditions

Originating condition arguments are among the most commonly used by contemporary empirically informed normative ethicists. The examples discussed in this section are intended to be illustrative of the style of argument, and this is certainly not an exhaustive survey.

For a first example: Many recent studies appear to show that our moral intuitions are highly sensitive to emotional manipulation, via environmental cues like physical cleanliness or emotional levity (Schnall et al. 2008; Valdesolo and DeSteno 2006; see also Wheatley and Haidt 2005; Eskine et al. 2011; Strohminger et al. 2011). Haidt (2001; Haidt and Bjorklund 2008) suggests that these correlations show virtually all moral thinking to be "rationalization" rather than genuine reasoning.[20] Greene (2008) makes a more selective argument, claiming that deontological intuitions in particular are shown by neuroscience to be correlated to an unreliable sort of emotional processing.[21] Both of these arguments rely on the normative premise that *emotional* influence undermines the justificatory status of a moral intuition, something one might very well dispute (see Sherman 1990; Nussbaum 2003; Nichols 2004; Narvaez 2010). For the moment, the point is only this: Haidt and Greene have used psychology to identify the influence of emotion where one might not have suspected it, so if one *does* agree that emotional influence is a disqualifying originating condition, then one ought to pay close attention.

[20]It should be noted that in Haidt's later work (Haidt 2012), he downplays the centrality of emotion and instead focuses on *automated* cognition. For critical discussion of Haidt's view, see Pizarro and Bloom (2003); Kennett and Fine (2009); Liao (2011); Huebner (2011), among others.

[21]See Greene (Greene et al. 2001, 2004) for the psychological background. For further discussion, see the references given in parenthesis (Singer 2005; Nichols and Mallon 2006; Allman and Woodward 2008; Kamm 2009; Kahane and Shackel 2010; Kumar and Campbell 2012). Berker (2009) offers a particularly comprehensive criticism of the empirical and normative aspects of Greene's argument.

Another argument of this sort works slightly differently. Rather than tying intuitions to emotion, this argument instead aims to extract the computational *rules* guiding our intuitions. The substantial literature on *moral heuristics* is central here. According to this literature, moral judgment is like judgment in other domains: Many of our intuitions result from the rapid, unconscious application of decision rules. These rules are usually reliable within their target domains, but under certain circumstances, they produce systematic errors.[22] Importantly, the conditions under which these errors arise are often extremely subtle, likely to be detected only by psychological techniques, rather than introspective self-monitoring. Of course, the claim that certain applications of the rules count as *errors* is a normative premise, not provided by psychology itself. But, again, psychology's contribution to this argument is irreplaceable; the reevaluations urged by the argument could not have occurred without empirical discoveries.[23]

All that said, it is worth drawing a cautionary line around these arguments. Their proponents often have stridently ambitious aims, attempting to undermine or displace long-established moral theories. Sometimes these revisionary ambitions appear to get the better of their authors, leading to minimization of the role of normative theory itself in their arguments. For example: Horowitz (1998, p. 381) and Greene (2008, p. 70) each spend a good deal of time demonstrating the psychological correlates of particular intuitions, then rather hastily assert that the correlated factors are "nonmoral" or "morally irrelevant," with little to no justification for these assertions. Many readers go away unconvinced. But the argument of this section does not depend on the specifics of any of these examples – they are meant simply to illustrate the form of originating conditions arguments. Certainly more work needs to be done on the normative premises of these arguments – but the logic itself seems clear.[24]

---

[22]For more on moral heuristics, see the references given in parenthesis (Baron 1994; Horowitz 1998; van Roojen 1999; Kamm 1998; Sunstein 2005; Gigerenzer 2008; Appiah 2008). Sinnott-Armstrong (2008) makes a related, and still more ambitious, argument, claiming that the influence of morally irrelevant factors like framing effects or presentational order is so pervasive as to cast an epistemic shadow over *all* moral intuitions. This sort of argument is not unique to the moral domain – see Weinberg (2007); Alexander (2012, Chap. 4) for related criticism of the role of intuition in all areas of philosophy.

[23]Many responses to the moral heuristics argument and Greene's argument, including most of those mentioned in footnotes above, involve challenging their particular empirical or normative premises. Another response, suggested by Levy (2006), claims that a constructivist meta-ethic can immunize us to many of these challenges. If one holds that morality simply *is* whatever our intuitions (or some suitably restricted class of them) correspond to, then there is little danger of our intuitions turning out driven by "morally irrelevant" factors – the psychological findings will simply tell us what our moral beliefs have been committed to all along.

[24]See Rini (forthcoming) for discussion of a broader theoretical framework for arguments of this sort. Similar moderate approaches – sympathetic to empirical investigation, but issuing cautionary qualifications – can be found in the references given in parenthesis (Stevenson 1944, p. 123; Baier 1985, p. 224; Noble 1989, p. 53; Held 1996, p. 83; Appiah 2008; Tiberius 2010; Kahane forthcoming). But see Machery (2010) for a skeptical argument, accusing the moderate approach of circularity.

Conclusion

This chapter has identified three central aims of normative ethics, shown how each aim might seem to suggest strong limitations on the relevance of empirical moral psychology, then argued that in fact, each aim actually supports an important role for psychological inquiry. Each of these discussions thus constitutes an independent argument for sustaining and expanding normative ethicists' engagement with empirical findings.

But these disciplinary interactions are at quite early stages. The arguments discussed above only hint at the rich range of questions still to be investigated. Will new psychological discoveries prove more difficult to accommodate to reflective, practical agency? If our psychology imposes limits on what morality can demand of us, should we try to change our psychology? How pervasive are undermining originating conditions – will recognizing their extent risk general moral skepticism? We do not have answers to these questions. We do not yet even know precisely how to ask these questions in a way that they might be answered. Philosophers and psychologists have a lot of work to do together.[25]

Cross-References

- ▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
- ▶ Brain Research on Morality and Cognition
- ▶ Consciousness and Agency
- ▶ Moral Intuition in Philosophy and Psychology
- ▶ No Excuses: Performance Mistakes in Morality
- ▶ The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy
- ▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering

References

Alexander, J. (2012). *Experimental philosophy: An introduction* (1st ed.). Cambridge, UK: Polity.

Allman, J., & Woodward, J. (2008). What are moral intuitions and why should we care about them? A neurobiological perspective. *Philosophical Issues, 18*(1), 164–185. doi:10.1111/j.1533-6077.2008.00143.x.

[25]The contents of this chapter benefited significantly from discussions with Tommaso Bruni, Nora Heinzelmann, Guy Kahane, and Felix Schirmann, and from written comments by Stephan Schleim and an anonymous referee. This research was part of the project "Intuition and Emotion in Moral Decision Making," funded by the VolkswagenStiftung's European Platform for Life Sciences, Mind Sciences, and the Humanities (grant II/85 063).

Annas, J. (2005). Comments on john Doris's lack of character. *Philosophy and Phenomenological Research, 71*(3), 636–642.
Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy, 33*(124), 1–19.
Appiah, K. A. (2008). *Experiments in ethics*. Cambridge: Harvard University Press.
Baier, A. (1985). Theory and reflective practices. In *Postures of the mind*. Minneapolis: Univ of Minnesota Press.
Baron, J. (1994). Nonconsequentialist decisions. *The Behavioral and Brain Sciences, 17*(01), 1–10. doi:10.1017/S0140525X0003301X.
Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs, 37*(4), 293–329. doi:10.1111/j.1088-4963.2009.01164.x.
Casebeer, W. D., & Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy, 18*(1), 169–194.
Churchland, P. (2000). Rules, know-how, and the future of moral cognition. *Canadian Journal of Philosophy, 30*(Supplement), 291–306.
Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton: Princeton University Press.
Clark, A. (1996). Connectionism, moral cognition, and collaborative problem solving. In L. May, M. Friedman, & A. Clark (Eds.), *Minds and morals: Essays on cognitive science and ethics* (pp. 109–127). Cambridge: MIT Press.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. doi:10.1016/j.cognition.2008.03.006.
Cushman, F., & Young, L. (2009). The psychology of dilemmas and the philosophy of morality. *Ethical Theory and Moral Practice, 12*, 9–24.
Daniels, N. (1980). On some methods of ethics and linguistics. *Philosophical Studies, 37*(1), 21–36. doi:10.1007/BF00353498.
Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*, 100–108.
De Waal, F. (2006). *Primates and philosophers: How morality evolved*. Princeton: Princeton University Press.
Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge, UK: Cambridge University Press.
Doris, J. M. (2009). Skepticism about persons. *Philosophical Issues, 19*(1), 57–91.
Doris, J. M., & Stich, S. (2007). As a matter of fact: Empirical perspectives on ethics. In F. Jackson & M. Smith (Eds.), *The Oxford handbook of contemporary philosophy* (1st ed., Vol. 1, pp. 114–153). Oxford: Oxford University Press. Retrieved from http://www.oxfordhandbooks.com/oso/public/content/oho_philosophy/9780199234769/oxfordhb-9780199234769-chapter-5.html
Dupoux, E., & Jacob, P. (2007). Universal moral grammar: A critical appraisal. *Trends in Cognitive Sciences, 11*(9), 373–378. doi:10.1016/j.tics.2007.07.001.
Dworkin, R. (1996). Objectivity and truth: You'd better believe it. *Philosophy & Public Affairs, 25*(2), 87–139.
Dwyer, S. (2006). How good is the linguistic analogy? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Culture and cognition*. Oxford: Oxford University Press.
Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science, 22*(3), 295–299. doi:10.1177/0956797611398497.
Flanagan, O. (1993). *Varieties of moral personality: Ethics and psychological realism*. Cambridge: Harvard University Press.
Flanagan, O. (1996). Ethics naturalized: Ethics as human ecology. In L. May, M. Friedman, & A. Clark (Eds.), *Minds and morals: Essays on cognitive science and ethics* (pp. 19–43). Cambridge: MIT Press.

Foot, P. (2003). *Natural goodness*. Oxford: Oxford University Press.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.

Fried, C. (1978). Biology and ethics: Normative implications. In *Morality as a biological phenomenon: The presuppositions of sociobiological research* (pp. 187–197). Berkeley: University of California Press.

Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology* (The cognitive science of morality: Intuition and diversity, Vol. 2, pp. 1–26). Cambridge: MIT Press.

Gilligan, C. (1982). *In a different voice: Psychology theory and women's development*. Cambridge: Harvard University Press.

Goodpaster, K. E. (1982). Kohlbergian theory: A philosophical counterinvitation. *Ethics, 92*(3), 491–498.

Graham, P. A. (2011). "Ought" and ability. *Philosophical Review, 120*(3), 337–382. doi:10.1215/00318108-1263674.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (The neuroscience of morality: Emotion, brain disorders, and development, Vol. 3, pp. 35–80). Cambridge: MIT Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. doi:10.1126/science.1062872.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400. doi:10.1016/j.neuron.2004.09.027.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion* (1st ed.). New York: Pantheon.

Haidt, J., & Bjorklund, F. (2008). Social intuitions answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (The cognitive science of morality: Intuition and diversity, Vol. 2, pp. 181–218). Cambridge: MIT Press.

Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). A study of prisoners and guards in a simulated prison. *Naval Research Review, 30*, 4–17.

Hare, R. M. (1952). *The language of morals*. New York: Oxford University Press.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society, 99*, 315–331.

Hauser, M. D. (2006). The liver and the moral organ. *Social Cognitive and Affective Neuroscience, 1*(3), 214–220. doi:10.1093/scan/nsl026.

Held, V. (1996). Whose agenda? Ethics versus cognitive science. In L. May, M. Friedman, & A. Clark (Eds.), *Minds and morals: Essays on cognitive science and ethics* (pp. 69–88). Cambridge: MIT Press.

Hooker, B. (1990). Rule-consequentialism. *Mind, 99*(393), 67–77.

Horgan, T., & Timmons, M. (2009). What does the frame problem tell us about moral normativity? *Ethical Theory and Moral Practice, 12*(1), 25–51.

Horowitz, T. (1998). Philosophical intuitions and psychological theory. *Ethics, 108*(2), 367–385.

Huebner, B. (2011). Critiquing empirical moral psychology. *Philosophy of the Social Sciences, 41*(1), 50–83. doi:10.1177/0048393110388888.

Hume, D. (1739). A treatise of human nature. London: Penguin.

Hursthouse, R. (2002). *On virtue ethics*. Oxford: Oxford University Press.

Johnson, M. L. (1996). How moral psychology changes moral theory. In L. May, M. Friedman, & A. Clark (Eds.), *Minds and morals: Essays on cognitive science and ethics* (pp. 45–67). Cambridge: MIT Press.

Joyce, R. (2006). *The evolution of morality* (1st ed.). Cambridge: MIT Press.

Kahane, G. (2013). The armchair and the trolley: An argument for experimental ethics. *Philosophical Studies, 162*(2), 421–445.
Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind and Language, 25*(5), 561–582. doi:10.1111/j.1468-0017.2010.01401.x.
Kamm, F. M. (1998). Moral intuitions, cognitive psychology, and the harming-versus-not-aiding distinction. *Ethics, 108*(3), 463–488.
Kamm, F. M. (2009). Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs, 37*(4), 330–345. doi:10.1111/j.1088-4963.2009.01165.x.
Kamtekar, R. (2004). Situationism and virtue ethics on the content of our character. *Ethics, 114*(3), 458–491.
Kant, I. (1785). *Groundwork for the metaphysics of morals*. (trans: Wood, A.W.). New Haven: Yale University Press.
Kennett, J., & Fine, C. (2009). Will the real moral judgment please stand up? The implications of social intuitionist models of cognition for meta-ethics and moral psychology. *Ethical Theory and Moral Practice, 12*(1), 77–96.
Kitcher, P. (2011). *The ethical project*. Cambridge: Harvard University Press.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(279), 190–194. doi:10.1111/1467-8284.00419.
Knobe, J. (2010). Person as scientist, person as moralist. *The Behavioral and Brain Sciences, 33*(4), 315–329. doi:10.1017/S0140525X10000907. discussion 329–365.
Kohlberg, L. (1971). From "is" to "ought": How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.), *Cognitive development and epistemology*. New York: Academic.
Korsgaard, C. M. (1996a). Creating the kingdom of ends: Reciprocity and responsibility in personal relations. In *Creating the kingdom of ends* (pp. 188–223). Cambridge University Press.
Korsgaard, C. M. (1996b). *The sources of normativity*. Cambridge: Cambridge University Press.
Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology, 25*(3), 311–330.
Levy, N. (2006). Cognitive scientific challenges to morality. *Philosophical Psychology, 19*(5), 567–587. doi:10.1080/09515080600901863.
Levy, N. (2009). Empirically informed moral theory: A sketch of the landscape. *Ethical Theory and Moral Practice, 12*(1), 3–8. doi:10.1007/s10677-008-9146-2.
Liao, S. M. (2011). Bias and reasoning: Haidt's theory of moral judgment. In T. Brooks (Ed.), *New waves in ethics*. Basingstoke: Palgrave Macmillan.
Machery, E. (2010). The bleak implications of moral psychology. *Neuroethics, 3*(3), 223–231. doi:10.1007/s12152-010-9063-7.
MacIntyre, A. C. (1981). *After virtue: A study in moral theory*. South Bend: Notre Dame Press.
Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. London: Penguin.
Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment* (3rd ed.). Cambridge: Cambridge University Press.
Milgram, S. (1973). *Obedience to authority*. New York: Harper Torchbooks.
Moody-Adams, M. (2002). *Fieldwork in familiar places: Morality, culture, and philosophy* (Newth ed.). Cambridge: Harvard University Press.
Moore, G. E. (1903). *Principia ethica*. Cambridge: Cambridge University Press.
Nagel, T. (1978). Ethics as an autonomous theoretical subject. In G. S. Stent (Ed.), *Morality as a biological phenomenon: The presuppositions of sociobiological research* (pp. 198–205). Berkeley: University of California Press.
Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press.
Nagel, T. (1997). *The last word*. Oxford: Oxford University Press.
Narvaez, D. (2010). The emotional foundations of high moral intelligence. *New Directions for Child and Adolescent Development, 2010*(129), 77–94. doi:10.1002/cd.276.
Narvaez, D., & Lapsley, D. K. (Eds.). (2009). *Personality, identity, and character: Explorations in moral psychology* (1st ed.). Cambridge: Cambridge University Press.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford: Oxford University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*(3), 530–542. doi:10.1016/j.cognition.2005.07.005.

Noble, C. N. (1989). Normative ethical theories. In S. G. Clark & E. Simpson (Eds.), *Anti-theory in ethics and moral conservatism* (pp. 49–64). Albany: SUNY Press.

Nussbaum, M. C. (2003). *Upheavals of thought: The intelligence of emotions*. Cambridge: Cambridge University Press.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*(3), 145–171. doi:10.1016/0162-3095(96)00041-6.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*(3), 467–478. doi:10.1037/0022-3514.64.3.467.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: Comment on haidt (2001). *Psychological Review, 110*(1), 193–196. discussion 197–198.

Prinz, J. J. (2008). Resisting the linguistic analogy. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (The cognitive science of morality: Intuition and diversity, Vol. 2, pp. 157–170). Cambridge: MIT Press.

Rawls, J. (1951). Outline of a decision procedure for ethics. *Philosophical Review, 60*(2), 177–197.

Rawls, J. (1971). *A theory of justice* (1st ed.). Cambridge: Harvard University Press.

Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association, 48*, 5–22. doi:10.2307/3129858.

Rawls, J. (1987). The idea of an overlapping consensus. *Oxford Journal of Legal Studies, 7*(1), 1–25.

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. (1999). *Postconventional moral thinking: A Neo-kohlbergian approach*. Mahwah: Psychology Press.

Rini, R. A. (2011). *Within is the fountain of good: Moral philosophy and the science of the nonconscious mind*. Doctoral Thesis. New York: New York University.

Rini, R. A. (unpublished manuscript). *Kantian autonomy and piagetian autonomy*.

Rini, R. A. (2013). Making psychology normatively significant. *The Journal of Ethics, 17*(3), 257–274

Ross, L., & Nisbett, R. E. (1991). *The person and the situation* (70th ed.). New York: Mcgraw-Hill College.

Sabini, J., & Silver, M. (1982). *Moralities of everyday life*. New York: Oxford University Press.

Sabini, J., & Silver, M. (2005). Lack of character? Situationism critiqued. *Ethics, 115*(3), 535–562.

Scanlon, T. M. (2002). Rawls on justification. In S. Freeman (Ed.), *The Cambridge companion to Rawls* (pp. 139–167). Oxford: Oxford University Press.

Schleim, S., & Schirmann, F. (2011). Philosophical implications and multidisciplinary challenges of moral physiology. *Trames, 15*(2), 127–146.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality & social psychology bulletin, 34*(8), 1096–1109. doi:10.1177/0146167208317771.

Sherman, N. (1990). The place of emotions in Kantian morality. In O. Flanagan & A. Rorty (Eds.), *Identity, character and morality: Essays in moral psychology*. Cambridge: MIT Press.

Sidgwick, H. (1907). *The methods of ethics*. Hackett.

Singer, P. (1981). *The expanding circle: Ethics and sociobiology*. Oxford: Oxford University Press.

Singer, P. (2005). Ethics and intuitions. *Journal of Ethics, 9*(3–4), 331–352.

Sinnott-Armstrong, W. (2008). Framing moral intuition. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol 2. The cognitive science of morality: Intuition and diversity* (pp. 47–76). Cambridge: MIT Press.

Sripada, C. S. (2005). Punishment and the strategic structure of moral systems. *Biology and Philosophy, 20*, 767–789.

Stevenson, C. L. (1944). *Ethics and language*. New Haven: Yale University Press.
Stich, S. (1993). Moral philosophy and mental representation. In M. Hechter, L. Nadel, & R. Michod (Eds.), *The origin of values* (pp. 215–228). New York: Adine de Gruyer.
Strawson, P. F. (1974). Freedom and resentment. In his *Freedom and resentment and other essays*. Routledge:London: Methuen and Co. pp. 1–25.
Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition, 119*(2), 295–300. doi:10.1016/j.cognition.2010.12.012.
Sunstein, C. R. (2005). Moral heuristics. *The Behavioral and Brain Sciences, 28*(4), 531–542.
Tiberius, V. (2010). Appiah and the autonomy of ethics. *Neuroethics, 3*(3), 209–214. doi:10.1007/s12152-010-9064-6.
Turiel, E. (1983). *The development of social knowledge*. Cambridge: Cambridge University Press.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476–477. doi:10.1111/j.1467-9280.2006.01731.x.
Van Roojen, M. (1999). Reflective moral equilibrium and psychological theory. *Ethics, 109*(4), 846–857.
Weinberg, J. M. (2007). How to challenge intuitions empirically without risking skepticism. *Midwest Studies In Philosophy, 31*(1), 318–343. doi:10.1111/j.1475-4975.2007.00157.x.
Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*(10), 780–784. doi:10.1111/j.1467-9280.2005.01614.x.
Williams, B. (1985). *Ethics and the limits of philosophy*. Cambridge: Harvard University Press.
Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*(4), 1912–1920. doi:10.1016/j.neuroimage.2008.01.057.

11 Moral Intuition in Philosophy and Psychology

Antti Kauppinen

Contents

Abstract

Ethicists often appeal to moral intuitions in defending a theory. In this practice, the contents of intuitions are taken to support moral beliefs in a way that is often compared to the way the contents of perception support empirical beliefs. Philosophers have defended a variety of positions about the nature of such states. Intuitionists believe intuitions are either self-evident beliefs or intellectual appearances, while Coherentists think of them as considered judgments that need to be balanced against each other in a process of reflective equilibrium. Intuition skeptics reject either kind of justificatory role. Such skepticism has recently received support from psychological studies of moral intuition. In contrast to philosophers, psychologists typically think of intuitions as snap

A. Kauppinen
Trinity College, Dublin, Ireland
e-mail: KAUPPINA@tcd.ie

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_163,

judgments that result from automatic, nonconscious, and often affective processing. Some argue that they are likely to be responsive to morally irrelevant factors. Yet even if this is true of snap judgments, it is not clear what follows for the epistemic standing of the kind of states that philosophers talk about. The aim of this chapter is to clarify the various philosophical and psychological conceptions of moral intuition in order to bring them in closer contact and help researchers in different disciplines to avoid talking past each other. In the final section, a sentimentalist account of moral intuition that may offer some hope of reconciling the philosophical and psychological approaches is quickly sketched.

Introduction

Almost everyone agrees that ethical disputes cannot be resolved by appealing to empirical evidence alone. How, then, can we settle which moral principles to adopt or what to think of a particular moral issue? It is not uncommon for philosophers to appeal to *intuitions* as fundamental evidence or source of evidence. Is this practice justifiable, particularly in the light of recent findings in psychology and neuroscience?

To begin with, we have to know what we are talking about. The term "intuition" has many related uses in philosophy and psychology. Generally speaking, intuition contrasts with reasoning and has connotations with spontaneity and insight. But closer examination reveals that there are a number of different phenomena in this area, with the result that we must take great care to avoid talking past one another. My goal in this chapter is to clarify these distinctions and relate different conceptions in different disciplines to each other, in the hope of understanding better when and why we may rely on intuitions in ethical inquiry.

The most obvious distinction is between intuition as a kind of psychological *faculty* and intuition as a psychological *state*. In the first sense, intuition contrasts with reason and vision, for example. It is something we can make use of when forming beliefs. Some Early Modern Rationalists may have believed that we have such a special capacity to gain immediate access to abstract truths. But few if any contemporary philosophers believe in any such faculty, so this sense can safely be left aside here.

The crucial questions thus concern the nature and significance of *intuition as a psychological state*.[1] What is it for someone to have a moral intuition that it is wrong to kill one in order to save five, for example, and what does it matter if they do? For philosophers intuitions are primarily identified by their *putative epistemic and dialectical role*. In particular, intuitions are supposed to play a *foundational* or *quasi-foundational* role in nonempirical justification. They are justifiers that require no further proof or at least initially credible starting points of inquiry. As such, they are not based on theory, but rather constitute data that theories have to account for or explain away.

[1] Sometimes the term "intuition" is also used for the *proposition* that is the *content* of the psychological state. I will leave aside this use here. The adverb "intuitively" and the adjective "intuitive" are used much more broadly and shouldn't be taken to entail that anyone has an intuition (*pace* Cappelen 2012).

So the question is: what if any psychological states are fit to play such a justificatory role and why? Philosophers differ sharply. For *intuitionists*, intuitions are either intellectual appearances, attractions to assent to propositions, or beliefs that result from merely adequately understanding their content. For *coherentists*, intuitions are considered judgments that are inputs to a process of reflective equilibrium. For *intuition skeptics*, in contrast, there is nothing that plays the intuition role – so-called intuitions are just beliefs or inclinations to believe that have no special justificatory status (Williamson 2007; Cappelen 2012).

Psychologists and other empirical scientists, in turn, at least typically think of intuitions as beliefs that result from automatic, nonconscious, and nonrational psychological processing. In this chapter, such psychological processes will be labelled *intuitive processes*. The beliefs that result from them will be called *intuitions** to distinguish them from intuitions in the sense that philosophers use. It turns out to be very important to understanding and addressing philosophical claims made by psychologists and neuroscientists that they are not always talking about the same thing as philosophers when the latter use the word "intuition." Yet at the same time, psychological results may nevertheless be very important for understanding intuitions in the philosophers' sense.

This chapter will begin with an overview of intuitions as understood in empirical moral psychology and briefly examines some claims made on the basis of empirical findings. It will then outline the various philosophical conceptions. The final section looks at how the two pictures might be reconciled.

Intuitions in Empirical Moral Psychology

As noted, psychologists typically define intuitions as beliefs we acquire without (conscious) reasoning (i.e., intuitions*). For example, according to Jonathan Haidt, a moral intuition can be defined as

> the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion. (Haidt 2001, p. 818)

It is an empirical question which of our moral judgments are intuitions* in this sense and which judgments result from reasoning, for example. The primary goal of empirical moral psychology here is to find out to what extent intuitive processes *explain* why people judge as they do and what role they play in moral agency. But both psychologists themselves and experimental philosophers have also drawn normative or epistemic conclusions from these explanations.

Dual Process Theory

It has proven useful in psychology to make a distinction between two kinds of psychological process. Some processes are fast, effortless, nonconscious,

uncontrolled, and non-intentional; often associative or pattern matching and nonlinear; and frequently affective. They are also often contained within relatively self-contained mental "modules" such as visual perception that are innate and typically products of natural selection. These processes form what is often called *System 1* or the Intuitive System. In contrast, the processes that belong to *System 2* or the reasoning system are relatively slow, effortful, conscious and attention-demanding, controlled, intentional, serial, and potentially rational. The evidence in favor of this contrast between conscious and nonconscious thinking is by now overwhelming (see Wilson 2002; Kahneman 2011).

From a philosophical perspective, it is crucial that System 1 comprises of very *heterogeneous* processes. Some System 1 processes are likely to result in false beliefs, especially when triggered outside the context in which they are adaptive. The heuristics and biases approach (Kahneman et al. 1982) has extensively studied such predictably irrational processes. However, other System 1 processes are likely to produce true beliefs. For example, I understand an ordinary English sentence in a flash, as a result of nonconscious processing I can't control. Yet, of course, I *know* what it means. So sometimes the outputs of System 1 are likely to be false, and sometimes they are likely to be true. It all depends on which System 1 processes are at issue and what the circumstances are like. Bearing this point in mind is very important in thinking about the epistemic relevance of intuitive processes.

Intuitions and Explanation

Why do we make the moral judgments we do? One kind of answer to this explanatory question concerns the *proximal* causes of moral judgments – the psychological states and processes that cause moral judgments. They are responses to the agent's (perceived) situation. A further explanatory question concerns the origins of the proximate processes: why do certain situational features trigger certain responses? The answers to this question are *distal* explanations.

It is a signature claim of much recent empirical moral psychology that the proximal causes of many, indeed most, judgments are automatic and often affectively laden intuitive processes, so that many of our moral beliefs are intuitions* (see Pizarro, this volume). Briefly, we're often unable to articulate our reasons, emotion-related areas of the brain are activated when we judge, and we evaluate quickly, constantly, and without taxing working memory. As a rule, System 2 becomes involved only when there's a problem – when intuitions conflict or the social context requires explicit articulation (Haidt 2001). No one denies that *some* moral judgments result from reasoning – though reasoning often serves to rationalize preexisting intuitions* after the fact.

So intuitive processes appear to proximally explain many of our moral judgments. What is the *distal* explanation for the intuitions* that we have? Many contemporary psychologists appeal to the *evolutionary benefits* of intuitions*. For example, according to Haidt's Moral Foundations Theory (Haidt 2012), natural selection has favored the development of six innate modules that produce the

various affective responses underlying intuitions*. Roughly, the story is that groups whose members respond negatively to things like causing innocents to suffer and disrespecting those with high social status and positively to helping those in need or respecting the elders are likely to outcompete groups whose members do not have such responses. Consequently, the genes that program for the adaptive responses become more prevalent by group selection. Different cultures then fine-tune these innate responses in different ways.

For my purposes, the details of this broadly plausible account do not matter. The important questions arise from the assumptions that intuitions* are hardwired and evolutionary fitness-enhancing.

Are Intuitions* Unreliable?

In the psychological literature it is typical to assume that moral beliefs that result from explicit reasoning are epistemically unproblematic. To be sure, from a philosophical perspective this is highly contingent – after all, there is such a thing as bad reasoning and false premises! But what should we think of the epistemic status of intuitions*?

One popular claim is that intuitions* are unreliable, because they result from *affective* processes. The best-known such argument is made by Joshua Greene (e.g., Greene 2008). According to Greene, only *nonconsequentialist* beliefs result from intuitive processes, while consequentialist beliefs are the result of reasoning. On the basis of this, he constructs an *a posteriori* argument for consequentialism (cf. Kauppinen 2013):

1. Empirical investigation shows that nonconsequentialist moral intuitions* are proximately caused by intuitive emotional reactions.
2. Intuitive emotional reactions don't justify the beliefs they cause, because they are sensitive to morally irrelevant features.
3. So, empirical investigation undermines the justification of nonconsequentialist moral intuitions*.
4. Nonconsequentialist moral theory rests crucially on nonconsequentialist intuitions*.
5. So, nonconsequentialist moral theory is epistemically unsupported.

Greene's critics, such as Berker (2009), have pointed out that the second premise is not an empirical or neuroscientific one. But as Greene rightly counters, empirical evidence in favor of the first premise still does crucial work in the argument. Specifically, he believes that it shows nonconsequentialist intuitions* result from negative emotional responses towards using personal force to harm others (Greene et al. 2009). The distal explanation for this is roughly that such reactions facilitated peaceful coexistence in small groups in human prehistory. The evolved "point-and-shoot morality," however, is likely to misfire in modern conditions, so we shouldn't rely on intuitions* in moral thinking (cf. Singer 2005). Greene's openly normative assumption is that whether we use personal force in causing harm is morally irrelevant. As he notes, this is something it is hard for anyone to deny.

There are, however, at least three major problems with this line of argument. First, there are nonconsequentialists who disavow appeal to intuitions about cases, most notably Kantians (e.g., Wood 2011). Even if most people's nonconsequentialist beliefs were based on unreliable gut reactions, it doesn't follow that nonconsequentialist theory isn't true or unsupported, as long as there is some other kind of justification for it. Second, even if consequentialist beliefs result from reasoning, it doesn't follow that they are justified, unless the premises of that reasoning are themselves justified. And at least some of those premises appear to rely on intuitions, such as the intuition that it is better to save more rather than fewer people, at least other things being equal. Nonconsequentialists are free to turn the tables and say that this intuition is a part of an evolved point-and-shoot morality that sometimes gets it right and sometimes doesn't. If, on the other hand, being the product of evolution shows that a moral belief is unjustified – which many now consider a red herring (see, e.g., Kahane 2011) – then the consequentialist is hoist by his own petard.

Finally, many have called into question the specific interpretations of empirical data that Greene offers (e.g., C. Klein 2011). Even if intuitions* are results of an affective process, it has certainly not been shown that all morally relevant emotions are responsive to features that are uncontroversially morally irrelevant. For example, the sort of reactions that classical sentimentalists thought central to morality, such as resentment and gratitude, are responsive to features such as being used as a mere means for another's ends or exceeding expectations. Nothing so far shows they couldn't confer justification to beliefs (see below).

Can Intuitions* Be Reliable?

As I noted earlier, there are System 1 processes that are potentially sources of knowledge. If (some) intuitions* result from some such process, they will be trustworthy. I'll discuss two recent proposals to this effect.

According to what I'll call the *Expert Intuition View*, moral intuitions* *can* be reliable in just the same way as what are called expert intuitions are, provided we have the right kind of training. Allman and Woodward (2008) argue that moral intuitions* are the output of a species of affective social cognition that can be trained to be responsive to moral features by the same sort of implicit learning that teaches nurses to recognize which infants are sick, for example. In this kind of learning, the learner is exposed to cues she may not able to articulate, forms judgments, and receives independent feedback that tells her whether her judgments are on the right track or not – for example, a child's temperature returns to normal in response to treatment (G. Klein 1998). With enough experience, her System 1 delivers functional "intuitions" about what to do.

The chief, and in my view fatal, problem with the Expert Intuition View is that one of the necessary conditions for implicit learning, namely, immediate and unambiguous feedback (see Kahneman and Klein 2009), is missing in the case of morality. If I act on the mistaken intuition* that it's acceptable for me to break

a promise to a student in order to get a bit of rest, what is the equivalent of an infant's fever getting worse? Nothing. Even if moral mistakes reliably have bad consequences for *others*, no one thinks there is reliably a negative signal for the *agent*. People can persist with mistaken moral views quite easily, especially if surrounded by the like-minded.

A different approach to trustworthy intuitions* is provided by the Moral Grammar View (Dwyer 1999; Mikhail 2011; Hauser 2006). According to it, the process that yields moral intuitions* is nonconscious and automatic, but strictly rule-governed and computational rather than associative and affective. This innate moral competence is analogous to Chomsky's universal grammar. Again like linguistic understanding, the rules that govern this System 1 process are inaccessible to ordinary users, but can be articulated by experts, who deduce the existence of rules like the Doctrine of the Double Effect on the basis of observational data about judgments.

This is no place to evaluate Mikhail and Hauser's explanatory theory. Supposing it to be true, what can we conclude about the epistemic status of intuitions*? Mikhail talks freely about knowledge and competence and seems to assume that beliefs that result from exercising competence are somehow correct. But there's a crucial disanalogy to language here: moral judgments purport to represent how things are and are not in any sense "correct" just because they are entailed by *some* system of rules. An error theorist about ethics might happily endorse the Moral Grammar View as showing that we are hardwired to make systematically false moral judgments. So the epistemic status of the Doctrine of the Double Effect, for example, remains a mystery, if this explanation is correct.

Intuitions in Moral Philosophy

Although there is some controversy about how common appeals to intuition are in philosophy in general (Cappelen 2012), there is little doubt that they play a major role in contemporary normative ethics. Intuitions (or intuited propositions) about either particular cases or general principles are treated as presumptively valid starting points of ethical inquiry. As W.D. Ross put it, "the moral convictions of thoughtful and well-educated people are the data of ethics just as sense-perceptions are the data of a natural science" (Ross 1930/2002, p. 41). Normative theories typically aim to capture intuitions or else find some way of undermining the authority of a particular intuition. But views about the nature and importance of intuitions differ widely.

Self-Evidence Intuitionism

The classical intuitionist view is that certain moral truths are *self-evident*, much in the way that mathematical axioms are to those who understand them. (Sometimes the term "intuitionism" is also used for a related metaphysical thesis that moral properties are nonnatural, but this chapter will focus on the epistemological use.) Here's Ross:

> That an act *qua* fulfilling a promise, or *qua* effecting a just distribution of good . . . is *prima facie* right, is self-evident; not in the sense that it is evident from the beginning of our lives, or as soon as we attend to the proposition for the first time, but in the sense that when we have reached sufficient mental maturity and have given sufficient attention to the proposition it is evident without any need of proof, or of evidence beyond itself. It is evident just as a mathematical axiom . . . is evident (Ross 1930, pp. 29–30).

On a natural reading, for me to have the intuition that fulfilling a promise is *prima facie* right is for me to find the proposition evident simply on the basis of understanding it properly and attending to it. According to Robert Audi's contemporary reformulation, an intuition is a cognitive state whose content one doesn't infer from what one believes or any theory but which one forms on the basis of an adequate understanding of the intuited proposition (Audi 2004, pp. 33–36). Its content is a *self-evident* proposition. A proposition is self-evident "provided an adequate understanding of it is sufficient both for being justified in believing it and for knowing it if one believes it on the basis of that understanding" (Audi 2004, p. 49).

Audi thus considers intuitions as *beliefs* that are individuated by their distinctive *justification and aetiology*. Ernest Sosa's (2007) related rationalist view in general epistemology differs in that he considers an intuition to be an *attraction to assent* to a proposition rather than a belief. This is because we can have an intuition that *p* even if we know that *p* is false – for example, all the lemmas of a paradox are intuitive. On either picture, given that mere understanding suffices for knowing their content, intuitions are instances or sources of a priori knowledge. Given that their content is not tautological, they are sources of knowledge about *synthetic* truths.

Nonmoral examples of putatively self-evident propositions include *Nothing is both red and blue all over*, *No vixens are male*, and *The existence of great-grandchildren requires at least four generations of people*. Moral intuitionists differ on what kind of moral propositions are self-evident. For Sidgwick, only the most *fundamental moral principles* can lay claim to self-evidence. According to him, they are that "the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other" and that "as a rational being I am bound to aim at good generally" (Sidgwick 1907, p. 382). From these he infers the truth of a form of Utilitarianism.

Ross, in contrast, believes that several *mid-level moral principles* are self-evident. For example, he believes it is self-evident that we have (defeasible) moral reason to keep our word, match rewards to desert, be grateful to benefactors, to refrain from injuring others, and to benefit others. These are among considerations that always weigh in favor or against performing an action. An agent's overall or final duty in a particular case is a function of these *pro tanto* reasons. Finally, while most intuitionists agree that *verdicts about particular cases* cannot be self-evident, H. A. Prichard and moral particularists deny this. We can just see what the right thing to do in a particular situation is. Indeed, Prichard thought this is epistemically prior: "If I cannot see that I ought to pay this debt, I shall not be able to see that I ought to pay a debt" (Prichard 2002, p. 4).

Recent work by Audi, Shafer-Landau (2003), and others has dispelled many standard objections to the Self-Evidence View. There is no appeal to a special faculty of intuition, only ordinary understanding. A proposition can be self-evident even if some people who understand it don't regard it as self-evident, certain, or even true – after all, all that matters is that they *would be justified* in believing it on the basis of mere understanding. Nor need a self-evident proposition be obvious, as adequate understanding may take time and effort. Given the preceding, it is not surprising if people *disagree* about self-evident propositions – even if they adequately understand them, they may be led to deny them as a result of indoctrination, bad theory, or self-interest. A person suffering from such issues might sincerely believe that she has an intuition – but insofar as her belief or attraction isn't justifiable by mere understanding of the content, she is mistaken about its nature.

Nevertheless, challenges remain. Although epistemological intuitionism is logically independent of nonnaturalist moral metaphysics, they are de facto allies. After all, if moral facts were natural facts, why couldn't we come to know them the way we come to know other natural facts? But if moral facts are nonnatural and thus causally inefficacious, while intuitions are psychological states with causal histories governed by natural laws, it would have to be a fantastic cosmic coincidence for the contents of the intuitions to align with the nonnatural facts (Bedke 2009). Further, mere adequate understanding of the content is supposed to justify belief in the self-evident proposition. Sometimes this makes sense. What it is to understand the concept of a vixen is, at least in part, to know that it does not apply to males. So it is no surprise that merely understanding it suffices for knowing the *conceptual* or *analytic* truth that no vixens are male. But how can mere understanding reach to *synthetic* truths that are not about relations between concepts? That turns out to be very hard to account for (for recent attempts, see Jenkins 2008 and Bengson 2010).

Seeming-State Intuitionism

In general epistemology, it has recently become popular to think of intuitions as intellectual appearances or seemings (Bealer 2000; Chudnoff 2011). This view has adherents in moral epistemology as well, as Michael Huemer's definition of moral intuition shows:

> An intuition that p is a state of its seeming to one that p that is not dependent on inference from other beliefs and that results from thinking about p, as opposed to perceiving, remembering, or introspecting. (Huemer 2005, p. 102)

Huemer's definition has two parts: an intuition is (a) a seeming and (b) the result of merely thinking about the proposition. Seemings or appearances in general are *non-doxastic*, propositionally contentful states: it may seem to me that the stick in the water is bent, even though I do not *believe* that it is. They have a *presentational phenomenology*: when we have them, it's as if we're directly presented with their objects or truth-makers. Consequently, they are *compelling*: they attract assent to their content and appear to rationalize belief. Some such seemings are

perceptual, such as my visual experience of having a computer in front of me. But others are, according to seeming-state intuitionists, *intellectual*: they result from merely thinking about the proposition. They claim that merely thinking about killing one in order to save five can give rise to a non-doxastic, presentational, and compelling experience of moral wrongness.

Suppose that when I merely think about it, cheating on my spouse seems morally wrong to me. Does this justify my believing so? Seeming-state intuitionists tend to subscribe to a view about justification called epistemic liberalism (Bengson 2010). According to this view, we are justified in taking things to be as they appear to be, unless we have sufficient reason to doubt the appearances. Not every experience is a seeming in the specified sense, so the view doesn't license belief in just anything we dream or fantasize about. But when it comes to genuine seemings, we're not epistemically blameworthy for taking them at face value, other things being equal.

Walter Sinnott-Armstrong (2006) argues against this kind of view in ethics. According to him, empirical evidence shows that other things are not equal: the way things seem to us, morally speaking, is often biased by partiality, beset by peer disagreement, clouded by emotion, subject to framing effects, and influenced by disreputable sources like religion. Such factors, in general, give us reason to doubt appearances of a particular kind, so we can't take them at face value without some sort of further confirmation. But the need for further confirmation, Sinnott-Armstrong says, means that intuitions are unfit to play a foundational role in justification, so intuitionism is false.

Coherentism and Intuitions

What are often described as intuitions also play a crucial role in the best-known coherentist approach to moral epistemology, the method of *reflective equilibrium* (Rawls 1971; Daniels 1979). Here is how Rawls describes it:

> People have *considered judgments* at all levels of generality, from those about particular situations and institutions up through broad standards and first principles to formal and abstract conditions on moral conceptions. One tries to see how people would fit their various *convictions* into one coherent scheme, each considered judgment whatever its level having a certain *initial credibility*. By dropping and revising some, by reformulating and expanding others, one supposes that a systematic organization can be found (Rawls 1974, p. 8, my emphasis).

As I have highlighted, Rawls talks about considered judgments and convictions as the initially credible (but potentially revisable or dispensable) starting points of moral inquiry. In this lightweight sense, intuitions need not be any special sort of mental state or have any particular kind of aetiology. The emphasis on *considered* judgments rules out unreflective gut reactions, however.

While the deflationary aspect of Rawlsian intuitions has its attractions, it also raises an immediate epistemic question. Why should the mere fact that we believe something yield *any* initial credibility to the believed proposition? Precisely because the aetiology of the beliefs doesn't enter the picture, considered ideological

or self-serving judgments seem to start out with the same status as rational insights. Coherentists might respond by waging that such beliefs fall out in the process, but insofar as it is path dependent (i.e., the outcome depends on what the inputs are), there is no guarantee that the outputs of reflective equilibrium aren't systematically biased.

Coherentists might appeal to the notion of *wide* reflective equilibrium, in which psychological, sociological, and other empirical facts are brought into the balancing act. This might rule out beliefs with some intuitively problematic aetiologies, such as beliefs based on knee-jerk reactions or sensitive to the use of personal force (see the discussion of Greene above). But why? Because these causal histories typically result in beliefs that do not fit with the rest of our moral beliefs. This means that if our moral beliefs are distorted to begin with, widening the reflective equilibrium won't help. The ideologically brainwashed will regard the process we regard as brainwashing as conducive to true beliefs.

Perhaps the best response to worries about both seeming-state intuitionism and coherentism is provided by Mark van Roojen (forthcoming). He acknowledges that the kind of considerations Sinnott-Armstrong puts forward may mean that intuitions don't suffice to justify belief on their own. But when a proposition is *both* the content of a moral intuition *and* coheres with other intuitive propositions, belief in it will be justified. The initial credibility provided by intellectual appearance is, as it were, confirmed by coherence. Since reflective equilibrium is applied selectively only to appearance-based beliefs, low-quality inputs to the balancing process are filtered out at least to some degree.

Conclusion

How do the psychological and philosophical views of moral intuitions relate to each other? Could both disciplines learn something from each other? Before trying to answer these questions, here is a map of the different views of intuition that discussed here:

As Table 11.1 makes obvious, many intuitions* are not intuitions in the philosophical sense. This has several consequences for contemporary debates in empirically informed ethics or experimental philosophy.

First, as has been pointed out, not all System 1 processes are epistemically equal. One kind of intuitive process that is particularly likely to issue in false beliefs is quick gut reaction. But for philosophers, it is emphatically not the case that the truth of an intuited proposition supposed to be manifest in a quick flash. It is *immediate* only in the sense that its justification is not mediated by some further, itself justified belief. Coming to an adequate understanding of a proposition or thinking about a thought experiment is an effortful, System 2 process that may take time, often compared to what it takes to appreciate a work of art (e.g., McMahan 2013). As Audi (2004) puts it, an intuition can be a conclusion of *reflection*, although it can't be a conclusion of *inference*. This kind of reflection is not the sole privilege of philosophers, but it is one of the things they're trained to do. This gives a positive reason to regard philosophers' intuitions as superior.

Table 11.1 Views about moral intuition

	Type of mental state	*Aetiology*	*Claimed epistemic standing*
Psychology	Belief	System 1	Most consider dubious
Self-evidence intuitionism	Belief or attraction to assent	Mere adequate understanding of content, which also justifies belief	Constitutes or is a source of knowledge of nonnatural facts
Seeming-state intuitionism	Appearance/seeming/presentation	Merely thinking about the content	Justifies belief in the absence of reason to doubt
Coherentism	Belief/conviction	Reflection, consideration	Initial credibility

Second, it is often claimed that surveys or questionnaires reveal people's moral intuitions about particular cases so that by varying the cases and performing statistical analyses, we can discover what ordinary people's intuitions are sensitive to (Kahane 2013). But as John Bengson (2013) has pointed out, this is simply not a valid inference. Even if responses reflect intuitions*, it doesn't follow that the subjects have an intuition in any sense that interests philosophers. This at least limits the usefulness of survey studies and complicates any empirical study of intuitions.

Future Directions

Although not all intuitions* are intuitions in the philosophical sense, the latter are nevertheless the outcome of some intuitive process rather than reasoning or inference. Could empirical evidence about the nonconscious functioning of brain and mind help understand intuitions and their role? I believe it is going to be one important ingredient in the mix. This chapter will finish by briefly sketching the case for thinking of intuitions as *manifestations of moral sentiments* (see Kauppinen forthcoming).

Start with the observation that moral intuitions appear to have a *distinctive and diverse phenomenology*. When something *seems* wrong, it often *feels* wrong – even if one *believes* it is right. This suggests that intuitions are non-doxastic experiences, as seeming-state intuitionists think, but not the same kind of experience as other intuitions. Second, moral seemings can directly *motivate* us to act and react. This is clearest in cases like Huck Finn's lying to slave catchers: although he didn't *believe* that it was the right thing to do (it went against everything he was taught), it nevertheless *felt* like the right thing, and this was sufficient to move him to act. Third, note that apparently manipulating the subjects' emotions results in change in their intuitions (e.g., Valdesolo and di Stefano 2006), while it is unlikely to change the mathematical or other run-of-the-mill intellectual intuitions.

All these special features of moral intuitions are readily accounted for if they are emotions. Which emotions? Those that manifest moral sentiments. Sentiments are dispositions to feel, act, think, and believe (cf. Prinz 2007). The sentiment of liking one's dog manifests itself in delight on seeing the animal, sadness when it is ill, noticing when its favorite food is for sale, desire to buy the food, and so on. The sentiment of moral disapprobation towards cheating on one's spouse manifests itself in anger towards someone who cheats on his spouse, guilt when one thinks of doing it oneself, desire to refrain from acting on certain desires, and so on. In this context, Kauppinen (forthcoming b) has argued the emotional manifestations of the sentiment constitute moral appearances: the anger you feel towards the unfaithful husband *presents* his action as morally wrong and attracts you to *believe* that it is morally wrong, while also having a distinctive phenomenal feel and motivational role. Since these emotions constitute moral seemings, they confer defeasible initial credibility to their contents just as other seemings do.

But aren't emotions subject to general epistemic defeaters, due to their characteristic fickleness and partiality? In this respect, not all emotions are created equal. *Canonical moral sentiments* are felt from what Hume called "The Common Point of View" – roughly speaking, they result from a process of impartially empathizing with the hedonic states and reactive attitudes of those affected by the actual or hypothetical action. When my anger or guilt is based on this kind of sentiment, the generic reasons for doubting emotional appearances are absent – the emotional reactions aren't fickle, rash, partial, or ill-informed. Further, such responses are not sensitive to what everyone agrees are morally irrelevant features, such as mere physical distance, but rather to features like being treated as mere means or not receiving equal reward for equal contribution, which are plausibly morally relevant.

So let's go back to the questions we started out with. I have the moral intuition that it is wrong for a doctor to grab a random person off the street and take his vital organs to save five others. According to the view just sketched, this intuition is trustworthy when it consists in an emotional response that I have when I merely think about the case in the canonical way – not just understanding the proposition but also imagining myself in the shoes of those affected by the action. This suggestion is compatible with the well-supported empirical hypothesis that emotions play a crucial causal role in moral judging. Although the response itself results from a System 1 process, it may be preceded by conscious, System 2 effort to reflect on the situation and empathize with different perspectives. Given that the sentimental intuition presents the action as wrong and isn't subject to standard defeaters, it is fit to play at least a quasi-foundational role in moral justification. So it seems that if at least some moral intuitions consist in a sharply delimited kind of emotional response, it will be possible to go fairly far in reconciling the psychological and philosophical conceptions without skeptical consequences.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Human Brain Research and Ethics
▶ Moral Cognition: Introduction
▶ Psychology and the Aims of Normative Ethics

References

Allman, J., & Woodward, J. (2008). What are moral intuitions and why should we care about them? A neurobiological perspective. *Philosophical Issues, 18*, 164–185.
Audi, R. (2004). *The good in the right*. Princeton: Princeton University Press.
Bealer, G. (2000). A theory of the a priori. *Pacific Philosophical Quarterly, 81*, 1–30.
Bedke, M. (2009). Intuitive non-naturalism meets cosmic coincidence. *Pacific Philosophical Quarterly, 90*, 188–209.
Bengson, J. (2010). *The intellectual given*. Dissertation, University of Texas at Austin.
Bengson, J. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research, 86*(3), 495–532
Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs, 37*(4), 293–329.
Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.
Chudnoff, E. (2011). What intuitions are like. *Philosophy and Phenomenological Research, 82*(3), 625–654.
Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy, 76*(5), 256–282.
Dwyer, S. (1999). Moral competence. In K. Murasugi & R. Stainton (Eds.), *Philosophy and linguistics* (pp. 169–190). Boulder: Westview Press.
Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 3, pp. 35–80). Cambridge: MIT Press.
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Pantheon.
Hauser, M. (2006). *Moral minds*. New York: Harper Collins.
Huemer, M. (2005). *Ethical intuitionism*. New York: Palgrave MacMillan.
Jenkins, C. (2008). *Grounding concepts: An empirical basis for arithmetical knowledge*. Oxford: Oxford University Press.
Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526.
Kahane, G. (2011). Evolutionary debunking arguments. *Noûs, 45*(1), 103–125.
Kahane, G. (2013). The armchair and the trolley. *Philosophical Studies, 162*(2), 421–445.
Kahneman, D. (2011). *Thinking fast and slow*. New York: Macmillan.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
Kauppinen, A. (2013). Ethics and empirical psychology. In M. Christen, J. Fischer, M. Huppenbauer, C. Tanner & C. van Schaik (eds.) *Empirically informed ethics*. New york: Springer.

Kauppinen, A. (forthcoming). Intuition and belief in moral motivation. In Gunnar Björnsson et al. (eds.) *Motivational internalism*. Oxford: Oxford University Press.
Klein, G. (1998). *Sources of power. How people make decisions*. Cambridge: MIT Press.
Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics, 4*, 143–162.
Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge: Cambridge University Press.
Prichard, H. A. (2002). *Moral writings*. Ed. Jim MacAdam. Oxford: Oxford University Press.
Prinz, J. (2007). *The emotional construction of morals*. New York: Oxford University Press.
Rawls, J. (1971). *A theory of justice*. Cambridge: Harvard University Press.
Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association, 48*, 5–22.
van Roojen, M. (forthcoming), Moral intuitionism, experiments, and skeptical arguments. In A. Booth and D. Rowbottom (eds.) *Intuitions*. Oxford: Oxford University Press.
Ross, W. D. (1930/2002). *The right and the good*. Ed. Philip Stratton-Lake. Oxford: Oxford University Press.
Shafer-Landau, R. (2003). *Moral realism: A defence*. New York: Oxford University Press.
Sidgwick, H. (1907). *Methods of ethics* (7th ed.). London: MacMillan & Co.
Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics, 9*, 331–352.
Sinnott-Armstrong, W. (2006). Moral intuitionism meets empirical psychology. In T. Horgan & M. Timmons (Eds.), *Metaethics after Moore*. Oxford: Oxford University Press.
Sosa, E. (2007). *A virtue epistemology I: Apt belief and reflective knowledge*. New York: Oxford University Press.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476–477.
Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell.
Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.
Wood, A. (2011). Humanity as an end in itself. In D. Parfit (Ed.), *On what matters* (Vol. 2, pp. 58–82). Oxford: Oxford University Press.

The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy

12

Stephan Schleim

Contents

Abstract

The pioneering neuroscience of moral decisions studies implementing the moral dilemma task by Joshua Greene and colleagues stimulated interdisciplinary experimental research on moral cognition as well as a philosophical debate on its normative implications. This chapter emphasizes the influence these studies had and continue to have on many academic disciplines. It continues with a detailed analysis of both the traditional philosophical puzzle and the recent psychological puzzle that Greene and colleagues wanted to solve, with a special focus on the conceptual and experimental relation between the two puzzles. The analysis follows the fundamental logics essential for psychological experimentation that is also employed within cognitive neuroscience: the logics of defining a psychological construct, operationalizing it, formulating a hypothesis, applying it in an experiment, collecting data, and eventually interpreting them.

S. Schleim
Faculty of Behavioral and Social Sciences, Theory and History of Psychology, Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

Neurophilosophy, Munich Center for Neurosciences, Ludwig-Maximilians-University Munich, Munich, Germany
e-mail: s.schleim@rug.nl

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_164,

In this manner, this chapter exemplifies an analytical structure that can be applied to many other examples in experimental (neuro-) philosophy, here coined "The Experimental Neurophilosophy Cycle." This chapter eventually discusses how the empirical findings and their interpretation, respectively, are related back to the original philosophical and psychological puzzles and concludes with conceptual as well as experimental suggestions for further research on moral cognition.

Introduction

> Paradoxes such as this mean job security for philosophers. They've been debating them for decades but have been unable to come up with a logical reason why sometimes it's OK to kill one person to save five, whereas other times it's not. Now, an interdisciplinary team has offered its philosopher colleagues a helping hand. (Helmuth 2001, p.1971)

This excerpt from the report that accompanied the pioneering neuroscience of moral decisions study by Joshua Greene and colleagues in *Science* (Greene et al. 2001) has an implication that may be disconcerting for ethicists and moral philosophers: If it is true that the moral dilemmas that so far resisted clear philosophical solutions but provided job security for these scholars can now be explained by psychologists and neuroscientists, then do we still need ethicists and moral philosophers? Or, put more reluctantly, is it possible that we will need them less in the future than we did in the past? This chapter, broadly speaking, is about the provocative question whether and in how far the tasks carried out by philosophers are likely to be replaced by scholars from other disciplines, and it will discuss this on the basis of the said findings of moral neuroscience, moral physiology (Schleim 2008), neuroscience of ethics, or however one would like to call this field of research.

The debate concerning the relation between philosophy understood as a quest for conceptual clarification, logical truth, or normative conclusions and the empirical disciplines has a long tradition. With the separation of psychology around the turn of the previous century and more recently with advances in the cognitive neurosciences, scholars began to investigate questions concerning consciousness and mind independently from their philosophical roots and employing empirical methods such as experimentation, surveys, and systematic observation.

Both within and outside of philosophy, different views on this relation exist: Investigating language practices of the neurosciences, one can emphasize the importance of philosophically clarifying empirically essential concepts in order to avoid confusion (Bennett and Hacker 2003); considering the brain as the primary organ with which we think and thus philosophize, one can argue that our neurobiology essentially limits our possible experiences which in turn has implications for our philosophical theories (Northoff 2004). Others remark skeptically that thousands of years of "philosophical speculation" on, for example, the phenomenon of consciousness have not been very informative and suggest to leave philosophical issues aside in order to face and possibly solve the challenge empirically (Crick and Koch 1998).

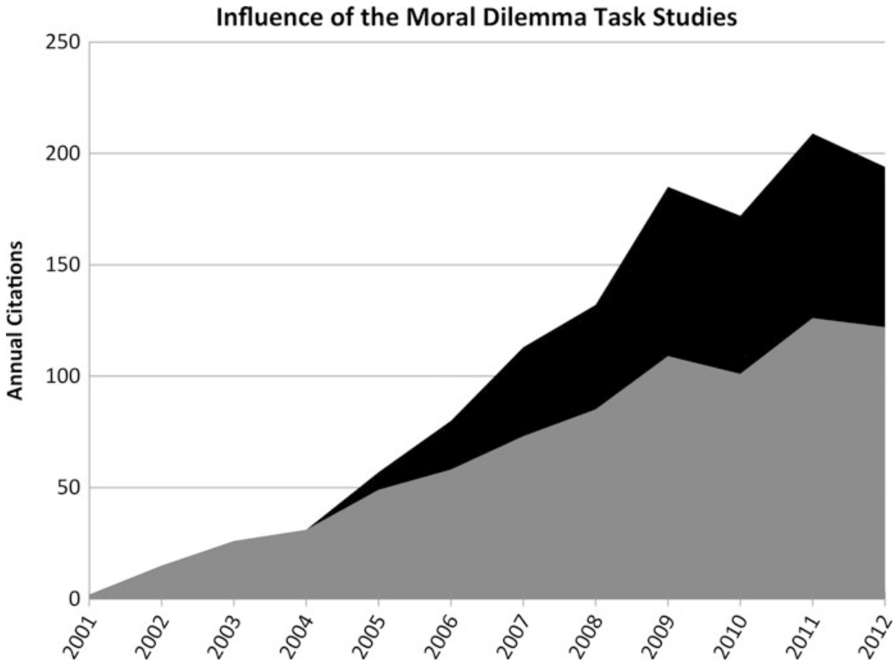

Fig. 12.1 Yearly citations of Greene and colleagues' moral dilemma task studies, published 2001 in Science (*gray*) and 2004 in Neuron (*black*). The graph shows a constant increase, totaling 1,216 for the whole period of 2001–2012 (Source: ISI Web of Science)

Particularly with regard to socially widely relevant issues concerning normative practices, such as the possibility of free will, moral responsibility, and the justification of the penal code, there even was a competition on whether empirical research or non-empirical argumentation has the final say (Libet 1985, and subsequent comments Roskies 2006a; Sie and Wouters 2010). Related to that domain, some philosophers used empirical cases to gain meta-ethical insight, such as in the plausibility of motivational internalism, the thesis that moral beliefs are intrinsically motivating (Levy 2007; Roskies 2006b). The probably most popular and until today widely cited example (see Fig. 12.1) used functional Magnetic Resonance Imaging (fMRI), a method to record brain signals in order to interpret them psychologically (see also ▶ Chap. 40, " Neuroimaging Neuroethics: Introduction"), not only to investigate why common people respond to moral dilemmas in the way they do, but also to explain why moral philosophers traditionally held incompatible views on these matters (Greene et al. 2001, 2004).

While others discussed the possibilities, prospects, and limitations of experimental philosophy generally (Alexander 2012; Knobe and Nichols 2008; *Philosophical Psychology* Special Issue 2010 Vol. 23 No. 3 & 4), this chapter deals with this popular investigation of moral dilemmas, its experimental preconditions, design, and psychological as well as philosophical interpretation. In this respect, the chapter is a case study in experimental (neuro-) philosophy, but its stepwise manner of analysis can – and

perhaps even should – be applied to other cases, too. Based on this analysis, the author will argue that, first, because this kind of experimentation presupposes certain conceptual definitions, it cannot replace philosophical conceptual clarification entirely; second, because it presupposes methodological conventions and decisions, its results are provisional; and third, the necessity of interpreting empirical findings with regard to research questions implies an amount of tentativeness. From this it follows that experimental (neuro-) philosophy – or empirical research in general – cannot replace philosophy but at best inform it and at worst confuse it. Based on this analysis, the author will also argue for the need of a critical philosophy of experimentation.

More concretely, the stepwise structure of the analysis consists in discussing and answering the following questions with respect to the moral dilemma task designed and applied by Joshua D. Greene and colleagues (Greene et al. 2001, 2004): First, what were the questions they wanted to answer? Second, how did they choose to answer the questions? Third, how did they answer the questions? Fourth, can the answers to the previous three questions be upheld after a critical discussion? That is, the analysis covers the fundamental logics essential for psychological experimentation that is also employed within cognitive neuroscience: the logics of defining a psychological construct, operationalizing it, formulating a hypothesis, applying it in an experiment, collecting data, and eventually interpreting them.

This chapter does not deal with the normative implications of such experimentation directly, an issue debated controversially elsewhere (Berker 2009; Greene and Cohen 2004; Greene 2008; Kahane 2013; Kauppinen, this section; Rini, this section; Singer 2005). Obviously, though, this analysis of the experiments' presumptions, findings, and interpretations is indirectly relevant to that discussion.

What Were the Questions They Wanted to Answer?

Greene and colleagues (Greene et al. 2001, 2004) derived their experimental idea from a famous philosophical puzzle, the moral question whether it is right (morally permissible or, stronger, perhaps even obligatory) or wrong (morally forbidden) to save the lives of a number of people when this requires an action that implies – more or less directly – the death of a smaller number of people, typically just one person (Foot 1978; Thomson 1986). Exemplary cases are the famous *trolley* (whether to divert a trolley by throwing a switch such that it will run over and kill one person instead of five), *footbridge* (whether to push a large man from a footbridge and thus killing him in order to stop a trolley that would otherwise run over and kill five persons), and *infanticide* (whether to smother one's crying baby to prevent enemy soldiers from finding and killing a group of hiding people, including the baby) dilemmas. Different answers to these questions are especially puzzling from a consequentialist's point of view that evaluates the moral status of an action exclusively with respect to the consequences – here: the number of people dead and alive – of an action; they might be less puzzling from a practical legal perspective, for example, where the first case rather resembles an instance of negligent homicide, the second manslaughter, and the third self-defense.

However, Greene and colleagues did not address the philosophical puzzle in the first place, but rather a psychological variant thereof: Why do many people commonly decide that it is right to sacrifice the few for the good of the many in some cases, particularly the trolley dilemma, but not others, particularly the footbridge dilemma? They proposed that different reactions may be due to people's emotional responses and capacities to deal with them cognitively, processes whose effects should be measurable in behavior and brain. For this reason, Greene and colleagues designed an experiment employing moral dilemmas of different kinds, the factor that is manipulated by the experimenter and called the *independent* variable, while recording reaction time and brain signals using fMRI, the *dependent* variables. In their second study (Greene et al. 2004), they addressed the more specific question which processes correlate with a particular kind of judgment, the one they called "utilitarian," when contrasted with "non-utilitarian" judgments. Since this classification has been contested (see below), "utilitarian," "consequentialist," and "deontological" will always be used in quotation marks in order to distinguish the experimental concepts from the more classical understanding proposed by moral philosophers.

In the long run, Greene and colleagues related their findings with regard to the psychological puzzle back to the original philosophical puzzle, particular the question what generally drives moral judgments of a deontological, that is, essentially duty-based, kind, as compared to moral judgments of a consequentialist, that is, exclusively consequence-based, kind (Greene et al. 2004; Greene 2008), a question that is discussed controversially until today (Paxton et al. in press; Kahane et al. 2012). In the course of this debate, the psychological answer was even used to speculatively explain why *philosophers* developed different moral accounts (Greene et al. 2004; Greene 2008). However, as stated in the introduction, the focus of this chapter is the psychological puzzle and how the researchers tried to answer it.

How Did They Choose to Answer the Questions?

Greene and colleagues introduced an experimental distinction that they thought reflected essential differences between trolley-type and footbridge-type cases, namely, the distinction between moral-*personal* and moral-*impersonal* dilemmas. For experiments measuring behavioral or brain responses, it is often insufficient to use just a few stimuli, in an extreme case only one, because the data tend to have a lot of variability, partially attributed to random processes called "noise." When repeating a measurement a sufficient number of times, in a sufficient number of subjects, the idea is that an average measure of the collected data should represent the "signal," that is the pattern truly related to the experimental conditions and not random fluctuations or other uncontrolled features. That is the reason why it would have been insufficient to just present the two or three original dilemmas, as they had been described and discussed before intensively by philosophers, to the experimental subjects, and general categories had to be developed.

The researchers stated that they considered their personal-impersonal-distinction for the experimental categories by no means definitive and primarily chose it for its usefulness, that is, for pragmatic reasons (Greene et al. 2001, 2004). However, this distinction was the central part of their operationalization to answer the psychological puzzle, because they expected that moral-personal dilemmas reflecting footbridge, where many people would not endorse the sacrifice, would be accompanied by signs of emotions in contrast to the moral-impersonal dilemmas reflecting trolley, where most would endorse the sacrifice. This distinction, as it was operationalized for the experiment, consisted of the following three requirements guaranteeing that the sacrifice in the moral-personal category would be brought about in an "up close and personal" manner (Greene et al. 2004; Greene 2008): First, the action would reasonably be expected to lead to serious bodily harm; second, this harm would not only consist in a deflection of a threat from one party to another, implying, in the researchers' opinion, a certain kind of agency; and third, this harm would be inflicted on a particular person or members of a particular group. Cases meeting these three criteria were considered as moral-personal dilemmas, otherwise as moral-impersonal.

The experimental methods chosen to provide the data to answer the psychological puzzle were an fMRI scanner, recording a primarily blood-oxygenation-related signal in the brain that is often considered a proxy for neural activation, though the correct interpretation of this signal is still a matter of basic research (Logothetis 2008; Schleim and Roiser 2009), in which the subjects, groups of circa ten to forty undergraduate students, were shown the dilemma texts, circa twenty per condition on a computer screen and had to use buttons to provide their answers; the time needed to provide these answers was also recorded. Note that because the experiment was carried out in slight variations for the publications 2001 and 2004, these figures represent approximations, which may be neglected for the purpose of this discussion.

For the more detailed question regarding the difference between "utilitarian" and "non-utilitarian" judgments, the researchers compared responses within a subclass of their stimulus material, namely, difficult moral-personal dilemmas, where "difficult" means that subjects needed more time for their answers compared to the other cases. When subjects endorsed the sacrifice action, Greene and colleagues considered this a "utilitarian" judgment, otherwise "non-utilitarian" (Greene et al. 2004), later discussed as "consequentialist" and "deontological" judgments, respectively (Greene 2008).

How Did They Answer the Questions?

Using a statistical model into which Greene and colleagues entered the respective experimental data (such as the different dilemma types and subjects), they calculated statistical maps that yielded significant differences in a number of brain areas previously associated with emotion, such as the medial frontal gyrus or posterior cingulate gyrus, or with working memory, such as the middle frontal gyrus or the parietal lobe. To understand these differences better, they carried out a further statistical analysis directly comparing the statistical values within the different areas for the moral-personal in contrast to the moral-impersonal condition (for a visualization of many

of these brain areas, please see the ► Chap. 9, "The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering" in this section). This demonstrated that the recorded brain signals within the emotion-related areas were relatively higher during the moral-personal than the moral-impersonal condition, and relatively lower in the working-memory-related areas. A similar statistical analysis for the response times showed that subjects, on average, took significantly longer to endorse a sacrifice in the moral-personal condition – almost seven seconds in contrast to only five when not endorsing the sacrifice – and that the difference in response times for the moral-impersonal condition was not significant (Greene et al. 2001; but see the discussion with respect to the re-analysis of the response time data by McGuire et al. 2009 below). In an analogous analysis for their second study, they could also relate both amygdalae, left and right, structures often associated with emotion, with the moral-personal condition (Greene et al. 2004).

For the more specific psychological puzzle, the brain signals related to "utilitarian" judgment, Greene and colleagues carried out so-called region-of-interest and whole-brain analyses. The former restricted the subsequent statistical tests to those areas that had previously been related to difficult moral-personal judgments compared to easy ones, that is, moral-personal cases for which the subjects needed more time to give their response; the latter tested for statistical differences within the whole brain, independent of previous analyses. In this manner, the "utilitarian" judgments were related to significantly higher brain signals in the posterior cingulate, superior/middle frontal gyrus, which they subsumed under the broad concept dorsolateral prefrontal cortex (DLPFC), precuneus, inferior parietal lobe, lingual gyrus, inferior parietal lobe, and inferior, middle, and superior temporal gyrus (Greene et al. 2004).

In the light of these findings, Greene and colleagues suggested that differences in emotional processing indeed provided the explanation for people's different moral judgments, and thus also the solution to the psychological puzzle: As they expected, reading moral-personal dilemmas and taking a judgment on them correlated with significantly higher activation in brain areas associated with emotion; the reaction time pattern suggested that those who endorsed the sacrifice of the few for the lives of the many had to overcome their emotional reaction, which took more time, in contrast to those who did not endorse these actions; and, finally, those who did so and thus took the "utilitarian" decision had significantly higher activation in the DLPFC, related to cognitive control (Greene et al. 2001, 2004).

This proposed solution to the psychological puzzle was then eventually related back to the original philosophical puzzle, also employing an evolutionary account of brain and psychological development: "We propose that the tension between the utilitarian and deontological perspectives in moral philosophy reflects a more fundamental tension arising from the structure of the human brain. The social-emotional responses that we've inherited from our primate ancestors [. . .] undergird the absolute prohibitions that are central to deontology. In contrast, the 'moral calculus' that defines utilitarianism is made possible by more recently evolved structures in the frontal lobes that support abstract thinking and high-level cognitive control" (Greene et al. 2004, p. 398). This was also interpreted to suggest that classical deontological moral philosophies such as the one proposed by Immanuel

Kant were actually not based in reason, as commonly believed, but instead in post hoc rationalized emotional responses (Greene et al. 2004; Greene 2008).

Discussion of the Answers to the Three Previous Questions

The previous sections covered the full “experimental neurophilosophy cycle” (see Fig. 12.2), starting out with the philosophical puzzle that inspired Greene and colleagues, how they translated this into a psychological puzzle, the psychological constructs and hypotheses they formulated, their operationalization, (some of) the data they collected in their experiments, the subsequent analysis/results, and eventually how they interpreted them, relating the data back to their original psychological and philosophical puzzle.

The illustration emphasizes the different steps in the cycle, with subsequent steps each depending on the previous ones. Figuratively speaking, just like the strength of a chain depends on its weakest link, so does the strength of the experimental neurophilosophy cycle depends on the weakest translational and methodological processes between the steps. For example, if the formulation of the psychological puzzle missed the essential philosophical points, then the research could still be psychologically intriguing, yet hardly be philosophically relevant. One central issue here is what kind of philosophical difference the moral-personal-impersonal-distinction reflects, the strategy Greene and colleagues used to provoke emotional reactions in their subjects and see how they affect their judgments. The experimenters conceded that this distinction was provisional, particularly chosen for pragmatical reasons (Greene et al. 2001, 2004). Indeed, Greene later explained that he no longer believed that the three criteria – serious personal harm that is not merely deflected – were adequate (Greene 2008). Yet, this operationalization was an essential precondition for the answer to the psychological puzzle that emotion drives judgments in the moral-personal cases like footbridge, but not in the moral-impersonal ones like trolley, and research trying to understand which morally salient psychological features are reflected in these stimulus categories continues (Waldmann et al. 2012). This critique need not affect the answer to the particular psychological puzzle concerning the “utilitarian” versus “non-utilitarian” distinction, though, since the analysis comparing these different kinds of judgments was confined to the (difficult) moral-personal cases only, and was thus independent from the contested personal/impersonal distinction.

Joshua Greene removes his psychological answer from the philosophical puzzle in his later work, understanding “consequential” and “deontological” judgments in a functional manner, indeed believing that “it is possible that philosophers do not necessarily know what consequentialism and deontology really are” (Greene 2008, p. 37). According to his functional understanding, it was particularly the disagreement between “consequentialists” and “deontologists” that allowed to distinguish them, namely, in such a manner that “consequentialistic” conclusions characteristically favored saving more lives at the costs of fewer lives, whereas “deontological” conclusions characteristically considered this as wrong despite the benefits.

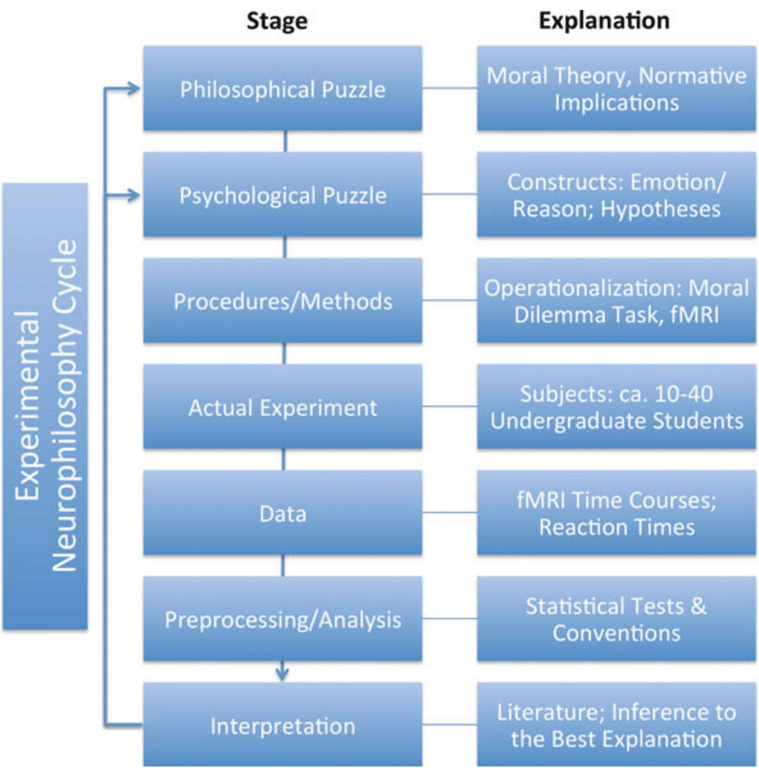

Fig. 12.2 The experimental neurophilosophy cycle: How Greene and colleagues translated their philosophical puzzle into a psychological one and finally relate their experimentally gathered data back to the original psychological and philosophical puzzles by means of their interpretation. This stepwise pattern of analysis can be applied to many examples in experimental (neuro-) philosophy

An experimenter can of course propose his or her own operationalizations, but the functional definition suggested in this case carries the risk of raising problems when linking the findings subject to that operationalization back to the original psychological and philosophical puzzle (Berker 2009; Kahane and Shackel 2010; Kamm 2009; Schleim and Schirmann 2011). That is, linking the psychological and the philosophical puzzle then carries the risk of committing an equivocation, a fallacy that while an argument looks consistent, it is in fact not because the meaning of concepts – here: "consequentialistic" and "deontological" – has essentially changed.

In their second study, Greene and colleagues proposed two psychologically necessary conditions for a judgment on "utilitarian" grounds, namely, abstract reasoning constituting a "utilitarian" analysis and the engagement of cognitive control enforcing the outcome of this analysis against competing, particularly emotional, pressures (Greene et al. 2004). They argued that they indeed found brain activities reflecting these psychological processes, but it may be doubted whether their presence alone is sufficient to identify the judgments as "utilitarian,"

given that a deontological analysis following Kant's categorical imperative literally, essentially includes the following: testing whether an action's maxim can be willed without contradiction to become a universal law (Kant 1785). This test certainly requires abstract reasoning, too, and favoring this action, for example, accepting the death of a dear one when not sacrificing a stranger to save his or her life, requires enforcing this outcome against competing, particularly emotional, pressures. Agreement on necessary and sufficient conditions of investigating judgments accurately reflecting influential moral theories is yet to be found, and it is thus not surprising that the controversial debate continues (Kahane et al. 2012; Paxton et al. in press).

The spatially localized brain activation found by Greene and colleagues cannot prove beyond reasonable doubt that subjects making a "utilitarian" judgment do so by employing abstract reasoning and cognitively controlling an emotional response. The reason for this is the general finding that the common psychological ontology – including central constructs such as "emotion" and "reason" – does not map clearly to brain areas, strictly associating one psychological construct with one brain area or one brain area with one psychological construct (Anderson 2010; Poldrack 2010). Meta-analyses of fMRI data show that all brain areas have a substantial amount of psychological diversity (Anderson and Pessoa 2011); even finding brain activation in a region like Broca's area in the prefrontal cortex that is classically considered to be highly functionally specific for language processing only adds relatively weak evidence to the interpretation that during a language task language processing was indeed occurring (Poldrack 2006). Accordingly, the presence of emotion processing cannot be inferred from statistically significant findings in a particular brain area (Pessoa 2008), not even the amygdalae (Pessoa and Adolphs 2010). At the current stage of knowledge, fMRI cannot be used as a "mind reading device" establishing the presence of a particular psychological process and perhaps it will never be (Logothetis 2008; Schleim and Roiser 2009).

This does not mean that these findings provide no information with regard to the psychological puzzles, but that a proper statistical account has to be employed that also considers, for example, experiments involving non-emotional tasks that correlated with the same brain patterns (Poldrack 2006). Greene's defense that his findings are compatible with a dual-process model encompassing both, quick emotional and slow deliberative approaches towards moral problems (Greene 2008, 2009; but see also Helion and Pizarro, this section; Kahane 2012), is in itself not very informative when these results may also be compatible with another one of the many other psycho-neurobiological accounts of moral cognition (Moll et al. 2005).

In addition, there is also contradicting evidence, such as the presence of relatively higher brain activation in areas such as the medial prefrontal cortex or the posterior cingulate during "utilitarian" judgment (see above; Greene et al. 2004; Moll and de Oliveira-Souza 2007) that the authors themselves related to emotional processing previously (Greene et al. 2001); and essential findings such as the emotional interference effect (see above; Greene et al. 2001) could not be replicated independently (Moore et al. 2008) and were related to poorly defined experimental stimuli in a re-analysis of the original data: When excluding the bad stimuli the response time

pattern originally suggesting the emotional interference disappeared, undermining the idea that the difference in judgments between moral-personal and impersonal dilemmas is driven by emotional influences (McGuire et al. 2009). These considerations do not call into question that experimental (neuro-) philosophy can be fruitful for the empirical sciences as well as philosophy, but suggest that it will be difficult and require both, conceptual clarification and further experimentation, in order to relate data to the respective psychological and philosophical puzzles.

Independent of the philosophical and psychological puzzles discussed in detail in this chapter, Greene himself speculated that the terms "deontology" and "consequentialism" refer to psychological natural kinds (Greene 2008). This suggests that while moral philosophers, in his view, may be wrong about the meaning of these moral theory concepts, empirical research might discover a natural essence allowing us to distinguish between these two (and possibly further) kinds of moral judgment similar to distinguishing between hydrogen and helium on the basis of their atomic number. Implications of this natural kinds approach for assumptions about the universality and innateness of moral judgment have been discussed elsewhere and would surpass the scope of this chapter (Mikhail 2011; Waldmann et al. 2012).

There is also a rich debate on the limited plausibility of the natural kinds concept within the human sciences in general, given the complex interaction between researcher and experimental subject in this domain (Gergen 2001; Hacking 1999). New evidence shows that even basic visual processes such as those underlying the Müller-Lyer illusion (that a line with two arrowheads pointing outwards looks shorter than a line of the same length with arrowheads pointing inwards) are culturally diverse and apparently mediated through learning (Henrich et al. 2010). In the related field of psychiatry that employs many psychological concepts like "attention" or "inhibition" it is widely conceded that no natural essences have been discovered so far despite strong efforts within the biomedical sciences, particularly genetics and neuroimaging (Kendler et al. 2011). Regardless of rather principal considerations on the possibility of psychological natural kinds, the current state of the art within psychology, cognitive neuroscience, and psychiatry as well as the cognitive diversity of the brain identified in fMRI meta-analyses described above make it appear very unlikely that even if such kinds existed, the currently available research methods are sufficient for discovering them.

Conclusion and Future Directions

While many inferences in the empirical sciences, in contrast to philosophical or mathematical logics, are rather inferences to the best explanation than definitive proofs, the previous critical discussion of several steps in the experimental (neuro-) philosophy cycle demonstrates that relating the data back to the original psychological and philosophical puzzles is subject to several views and decisions regarding the psychological constructs, hypotheses, procedures and methods, the analysis, and, eventually, the interpretation. Some of these views and decisions, for example, which model of moral cognition is most compatible with the available findings, can

be tested in further experimentation or through additional and more sophisticated analyses, but some of them, for example, how the constructs of emotion and reason should be understood, what the right method is for establishing their presence in a representative group of experimental subjects, or which statistical test provides the most reliable result, depend on conceptual aspects and convention within a scientific community.

While philosophy, just like any discipline, has to consider empirical findings inasmuch as it makes statements related to empirical subjects, empirical research, by contrast, presupposes theoretical stances on issues like the meaning of concepts which do not exclusively belong to the domain of philosophers, but for whose investigation philosophy provides helpful methods. This point is akin to the general account of the theory-ladenness of observation traditionally established within the philosophy of science (Duhem 1906; Kuhn 1962). However, it certainly deserves further research and discussion in which respect considerations regarding natural kinds differ between natural sciences in the narrow sense, like physics and chemistry, and the often interdisciplinary endeavors of human sciences and philosophy investigating multi-faceted subjects like consciousness, free will, or moral cognition.

Leaving these general issues aside, there is certainly much more to learn about the way people make moral judgments and engage, or fail to engage, in moral action. Whether or not this has implications for moral philosophy was not an issue in this chapter, but has been discussed controversially and extensively elsewhere (Berker 2009; Greene and Cohen 2004; Greene 2008; Kahane 2013; Kauppinen, this section; Rini, this section; Singer 2005). As the field of moral cognition continues to develop, there will certainly be much more discussion on this question. The studies by Greene and colleagues have certainly been inspiring for many neuroscientists, psychologists, philosophers, and other scholars alike, which is reflected in their citational success (see Fig. 12.1). The experimenters conceded that their approach was preliminary, which some might have neglected when relating their findings back to the original psychological and philosophical questions (see Fig. 12.2). Reaching agreement about the operationalization of moral judgments characteristic of important moral theories would help to integrate the findings of different studies with each other. Testing for differences and similarities between different groups of experts, students, and lay people might help as well, particularly when speculating about the way such individuals make their decisions; the author tried this for the case of legal experts such as lawyers and judges making moral and legal judgments (Schleim et al. 2011).

The recent success of and interest in experimental philosophy (Alexander 2012; Knobe and Nichols 2008; *Philosophical Psychology* Special Issue 2010 Vol. 23 No. 3 & 4) or "empirical neurophilosophy" (Northoff 2013), of which the studies by Greene and colleagues might actually be the most famous examples, is accompanied by a surge in critical methodological and theoretical publications questioning common ways of experimentation both within psychology and cognitive neuroscience (*Perspectives on Psychological Science* Special Issues Vol. 4 No. 3; 2010 Vol. 5 No. 6; 2012 Vol. 7 No. 6). Particularly with regard to fMRI, probably the most

successful method within human cognitive neuroscience throughout the last two decades (Friston 2009; Logothetis and Wandell 2004), the localizational where-question might not be as informative psychologically as often thought (Anderson 2010; Pessoa 2008; Poldrack 2006; Schleim and Roiser 2009). New methods yet to be developed will most likely provide better answers to what- and how-questions in the future in order to understand the correlations between brain signals and psychological constructs better, possibly allowing a better mapping between cognitive and brain ontology. For the time being, a critical philosophy of experimentation seems thus necessary in addition to experimental (neuro-) philosophy (Woolfolk 2013).

The initial statement by Helmuth that moral dilemmas provoking paradoxical reactions mean "job security" for philosophers accompanying the first study by Greene and colleagues suggests that once empirical researchers find a reasonable scientific explanation for people's – including philosophers' – moral judgments and views, philosophers might actually lose their jobs. At least in terms of philosophical inquiry, this chapter demonstrated that the experimental approach toward philosophically significant issues, by contrast, offers many new opportunities for scholars of various disciplines, including philosophers.

Acknowledgments The author would like to thank Professors Birnbacher, Gethmann, Hübner, Kahane, Kleingeld, Metzinger, Sellmaier, Stephan, and Walter as well as the Munich Neurophilosophy Group for the possibility to present earlier drafts of this work at their conferences or colloquia. The author would also like to acknowledge the helpful comments of the peer reviewers for clarifying some issues of this chapter. This paper was supported by the grant "Intuition and Emotion in Moral Decision-Making: Empirical Research and Normative Implications" by the Volkswagen Foundation, Az. II/85 063, and a generous travel grant by the Barbara Wengeler Foundation, Munich.

Cross-References

► Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
► Brain Research on Morality and Cognition
► Moral Intuition in Philosophy and Psychology
► Neuroscience, Neuroethics, and the Media

References

Alexander, J. (2012). *Experimental philosophy: An introduction*. Cambridge, UK/Malden, MA: Polity.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*(4), 245–266; Discussion 266–313.

Anderson, M. L., & Pessoa, L. (2011). *Quantifying the diversity of neural activations in individual brain regions*. Paper presented at the 33rd Annual Conference of the Cognitive Science Society, Austin, TX.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Malden: Blackwell.
Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs, 37*(4), 293–329.
Crick, F., & Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex, 8*(2), 97–107.
Duhem, P. P. M. (1906). *La théoriephysique: Son objet et sa structure*. Paris: Chevalier & Rivière.
Foot, P. (1978). *Virtues and vices and other essays in moral philosophy*. Berkeley: University of California Press.
Friston, K. J. (2009). Modalities, modes, and models in functional neuroimaging. *Science, 326*(5951), 399–403.
Gergen, K. J. (2001). Psychological science in a postmodern context. *Am Psychol*, 56(10), 803–813.
Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology. The neuroscience of morality: Emotion, brain disorders, and development* (Vol. 3, pp. 35–79). Cambridge, MA: MIT.
Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology, 45*(3), 581–584.
Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 359*(1451), 1775–1785.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400.
Hacking, I. (1999). The social construction of what? Cambridge, Mass: Harvard University Press.
Helmuth, L. (2001). Cognitive neuroscience. *Moral reasoning relies on emotion. Science*, 293(5537), 1971–1972.
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83; discussion 83–135.
Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind & Language, 27*(5), 519–545.
Kahane, G. (2013). The armchair and the trolley: An argument for experimental ethics. *Philosophical Studies, 162*(2), 421–445.
Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & Language, 25*(5), 561–582.
Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience, 7*(4), 393–402.
Kamm, F. M. (2009). Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs, 37*(4), 330–345.
Kant, I. (1785). *GrundlegungzurMetaphysik der Sitten*. Riga: J. F. Hartknoch.
Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychol Med*, *41*(6), 1143–1150.
Knobe, J. M., & Nichols, S. (2008). An experimental philosophy manifesto. In J. M. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 3–14). Oxford/New York: Oxford University Press.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Levy, N. (2007). The Responsibility of the psychopath revisited. *Philosophy, Psychiatry, & Psychology, 14*(2), 129–138.
Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), 529–539.
Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*(7197), 869–878.

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annual Review of Physiology, 66*, 735–769.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology, 45*(3), 577–580.

Mikhail, J. M. (2011). Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment. Cambridge, New York: Cambridge University Press.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences, 11*(8), 319–321.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*(10), 799–809.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*(6), 549–557.

Northoff, G. (2004). *Philosophy of the brain: The brain problem*. Netherlands: John Benjamins.

Northoff, G. (2013). Neurophilosophy. In C. G. Galizia & P.-M. Lledo (Eds.), *Neurosciences: From molecule to behavior* (pp. 75–80). Heidelberg/New York/Dordrecht/London: Springer.

Paxton, J. M., Bruni, T., & Greene, J. D. (in press). Are 'counter-intuitive' deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Social Cognitive and Affective Neuroscience* doi:10.1093/scan/nst102.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148–158.

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience, 11*(11), 773–783.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Poldrack, R. A. (2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed? *Perspectives on Psychological Science, 5*(6), 753–761.

Roskies, A. (2006a). Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences, 10*(9), 419–423.

Roskies, A. (2006b). Patients with ventromedial frontal damage have moral beliefs. *Philosophical Psychology, 19*(5), 617–627.

Schleim, S. (2008). Moral physiology, its limitations and philosophical implications. *Jahrbuchfür-Wissenschaft und Ethik, 13*, 51–80.

Schleim, S., & Roiser, J. P. (2009). FMRI in translation: The challenges facing real-world applications. *Frontiers in Human Neuroscience, 3*, 63.

Schleim, S., & Schirmann, F. (2011). Philosophical implications and multidisciplinary challenges of moral physiology. *Trames-Journal of the Humanities and Social Sciences, 15*(2), 127–146.

Schleim, S., Spranger, T. M., Erk, S., & Walter, H. (2011). From moral to legal judgment: The influence of normative context in lawyers and other academics. *Social Cognitive and Affective Neuroscience, 6*(1), 48–57.

Sie, M., & Wouters, A. (2010). The BCN challenge to compatibilist free will and personal responsibility. *Neuroethics, 3*(2), 121–133.

Singer, P. (2005). Ethics and intuitions. *Journal of Ethics, 9*, 331–352.

Thomson, J. J. (1986). *Rights, restitution, and risk: Essays, in moral theory*. Cambridge, Ma: Harvard University Press.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. Oxford: Oxford University Press.

Woolfolk, R. L. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy, 44*(1–2), 79–87.

Section III

Neuroscience, Free Will, and Responsibility

Neuroscience, Free Will, and Responsibility: The Current State of Play

13

Neil Levy

Contents

Abstract

A number of psychologists and neuroscientists have argued that experimental findings about the psychological basis of human behavior demonstrate that we lack free will, or that it is limited in ways we do not realize. In this chapter, I survey some of the main claims in this literature, with an eye to situating the contributions to come. I examine Benjamin Libet's work on the timing of awareness of action initiation, Daniel Wegner's claim that acting and thinking that one is acting dissociate, and related experimental work, and suggest that the threat to free will is smaller than has often been thought. I then turn to the question whether the brain is deterministic, and situate that question within philosophical debates. From global threats to free will, I move to local threats: the claim that the situationist literature in psychology shows that we lack free will under some circumstances. Finally, I examine recent developments in experimental philosophy, which aim to reveal ordinary people's views on free will.

N. Levy
Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, VIC, Australia
e-mail: nllevy@unimelb.edu.au; neil.levy@philosophy.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_15,

Introduction

Free will, and the closely related topic of moral responsibility, attracts more attention from outside philosophy than just about any other philosophical topic. In important part, this is due to the intrinsic significance of the topic: On many views, free will is required for moral and criminal responsibility, and therefore for the justification of large parts of the legal system (▶ Chap. 81, “Neurolaw: Introduction”). Perhaps for this reason, scientists from a variety of disciplines have found the topic irresistible, from physicists to evolutionary biologists (see, for instance, Suarez and Adams 2013). If free will exists, it is instantiated in human behavior, and the brain plays a central role in human behavior. For that reason, neuroscientists have been especially attracted to the question; correlatively, free will has become an important topic in neuroethics too. I will briefly discuss neuroscientific and, more broadly, cognitive scientific challenges to free will under the headings of *consciousness*, *determinism* and *situationism*, prior to a final discussion of the recent literature in *experimental philosophy* on free will.

Consciousness

One of the most famous experiments in all of neuroscience concerns free will. In ground-breaking work, Benjamin Libet and colleagues (1983) produced evidence that a brain event associated with voluntary movement, the *readiness potential*, begins to occur prior to subjects’ being aware that they have decided to act. This experiment has widely been interpreted as showing that people lack free will. People who advance this argument claim, roughly, that free will requires that we initiate our actions consciously, but if Libet is right we are not conscious of when we initiate action; rather, *first* we initiate action and only *subsequently* do we come to be aware of this fact. Our impression that we consciously initiated the action is an illusion, a confabulation in the jargon of neuropsychology.

We could buttress this argument with related work, both from neuroscience and from psychology. Consider, for instance, Soon et al. (2008): They found that patterns of activation in parietal and prefrontal cortex predicted which of two options subjects chose 7 whole seconds prior to the choice (the gap Libet found between readiness potential and awareness, by contrast, was less than half a second), or Daniel Wegner’s (2002) important work showing a double dissociation between thinking that one is acting and actually acting: Subjects sometimes think they have acted when they have not, and sometimes think they have not acted when they have. This last finding seems to show that our consciousness of our actions is an unreliable guide to the reality. Like Libet, Wegner interprets these results as showing that the conscious self actually plays a relatively small role in behavior: Rather, unconscious mechanisms do the actual work.

As Bayne and Pacherie, in their contribution, and Mele, in his, point out in this volume, this experimental work has been subjected to criticism which highlights

significant methodological problems. Mele (2009) in particular has been a very powerful and persuasive critic of Libet, and of more recent work building on his results. Readers interested in the methodological problems with this research will find details in those contributions. Here I want to focus on some philosophical questions that have not received the same degree of attention. We would do well to question the assumptions made by Libet, Wegner, and others who advance this argument against free will.

In particular, we should focus attention on the claim that if we initiate our actions unconsciously, we do not do so freely. Why think this is true? There is indeed evidence that consciousness is important to the ordinary conception of free will. In recent work, combining empirical data with philosophical analysis, for instance, Stillman et al. (2011) find "that people carefully considered what they should do before they engaged in actions they later regarded as good examples of free will" (391). But notice that the role for consciousness people apparently value is not the role that the scientists just mentioned claim it cannot play: Consciousness of *initiation* is quite different from conscious *deliberation*. Nothing in the empirical work mentioned threatens our belief that we act freely only when we act after careful deliberation (for more discussion, and argument that this kind of freedom can indeed be threatened by work in cognitive science, see Caruso 2012 and for a response see Levy 2014).

Determinism

Neuroscientists have also argued that the deterministic nature of brain processes rules out free will. The argument runs roughly as follows:

1. Free will requires that our actions be undetermined.
2. Neuroscience shows that the brain, which is the cause of our actions, is a deterministic causal machine.
3. Therefore, we lack free will.

As Mark Balaguer shows in his contribution to this volume, however, it is an open question whether the brain is indeed a deterministic system. It is compatible with what we know about both the brain and about physics that some or many brain processes are indeterministic: Certainly, neuroscientists cannot predict with high degrees of accuracy whether a particular neuron will fire or not (see Peter Tse 2013 for an argument that the brain is indeterministic and that therefore we possess free will).

Most philosophers do not think that the argument presented above advances the debate over free will in any useful way. That is because premise 1 of the argument begs the question against the majority view in philosophy regarding free will. *Compatibilists* deny that free will requires that our actions be undetermined. Free actions must not be coerced, compelled, or manipulated actions, they claim, but (deterministic) causation is none of these things. Indeed, some compatibilists have suggested that far from being a threat to freedom, deterministic causation is actually required for free will. A more common view is

that indeterminism does not contribute in any way to free will; it simply makes our actions subject to potentially responsibility-undermining luck (see Levy 2011 for discussion, and see Balaguer, this volume, for dissent). We will return to the issue of determinism shortly.

Situationism

Threats to free will from the sciences of the mind are not, as we have already seen, the exclusive province of neuroscience. A quite different set of experimental results seemed to many to represent a threat to free will: experiments showing that our actions can be significantly affected by features of the environment of which we are not aware, or of whose influence we are not aware. This is akin to the threat from lack of consciousness, but it is distinct: There is no special reason to think that agents are unaware of initiating their actions in these experiments, but there is good reason to think that they are unaware of the influence features of the situation have over them.

Much of this data comes from work on behavioral priming (now quite controversial in psychology, because of some high-profile failures to replicate well-known results). Consider for instance Williams and Bargh (2008). They found that subjects who held a warm drink judged another person as having a "warmer" personality than subjects who held a cold drink, and subjects holding a warm drink were more likely to choose a gift for a friend (rather than themselves) than subjects holding a cool drink (Williams and Bargh 2008). Subjects were aware of a stimulus – the temperature of the cup – but unaware of how it influenced their judgments.

This may seem like a prima facie threat to free will: People's willingness to give a gift to someone (and, as Amaya and Doris sketch in their contribution, to help another in need) is influenced by apparently irrelevant aspects of circumstances. We do not think that whether we act badly or well ought to be settled by these insignificant factors. Prima facie, the role these factors play seems to undermine our free will. Amaya and Doris argue that we should understand moral responsibility not as picking out a single monolithic concept but instead a family of related notions, and that some of these are applicable to agents in the circumstances of these experiments. If we understand free will (as most, though not all, philosophers do) as a power the exercise of which sometimes makes us morally responsible, we will find that the situationist data is no threat to free will after all.

I want to suggest a different but related reason why we might think that this data is not a special threat to free will: Though they certainly play a role in how we act, situational factors like these do not cause us to act contrary to our own values. The person who stops and helps the passer-by, but would have hurried on had they been under time pressure is neither a very decent nor a very bad person, I suggest: Rather, they are someone who genuinely has it in them to act well or badly. Either action is actually compatible with their values. Had they been really decent, the time pressure would not have been sufficient to make them pass by (only clearly excusing conditions would have done that), and had they been really bad, they would not have

helped under any circumstances. In some of the most famous experiments involving helping and harming behavior, such as Milgram's experiment involving the application of a shock to other people, most subjects went – shockingly – far. But some did not: Some resisted the situational pressure. Of course, this defense of free will against situationism concedes that freedom is subject to luck. Some people believe that luck is an independent threat to free will (I am one of them). My point is not that we can show that we are free despite the challenge: rather, it is that if there is a challenge here, it is not from social psychology at all.

Experimental Philosophy

Experimental philosophy is a relatively new movement that combines philosophical analysis with careful experimentation (▶ Chap. 12, "The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy"). Though experimental philosophers use a variety of methods and put them to a variety of purposes, one of the primary aims of the movement has been to probe the *intuitions* – roughly, the way things seem to be – of ordinary people. Analytic philosophy often treats intuitions as data (if someone in a particular thought experiment seems to lack knowledge, for instance, that is a good reason to think that someone in those circumstances would actually lack knowledge). It is often important to know whether one's own intuitions are idiosyncratic, or on the contrary widely shared. For that reason, experimental philosophers use carefully designed vignettes to see what factors influence the intuitions of ordinary people.

Under the heading of "determinism" above, I mentioned the compatibilist response to the argument that we lack freedom because the brain is (allegedly) a deterministic causal system. Some people have responded that compatibilism is so bizarre only a philosopher would believe it. Ordinary people, they claim, would need to be argued into accepting such a claim. Experimental philosophers have hotly debated whether compatibilism is highly unintuitive, or whether instead it strikes ordinary people as plausible. Of course, they do not ask ordinary people about "compatibilism." Rather they have given their subjects vignettes, describing the behavior of agents in worlds that are stipulated to be deterministic. They have then asked the subjects whether the agents are morally responsible for their actions (on the assumption, mentioned above, that if an agent is directly morally responsible for an action, the action was performed freely).

One fairly robust finding in the experimental literature is that ordinary people's attitudes depend in part upon how questions about free will are asked. In response to questions such as *if determinism is true, are people responsible for their actions*, a majority of ordinary people say *no*, just as the neuroscientists might have predicted. But when the question posed is more concrete in its form, asking whether a particular person is responsible for a particular crime in a determinist universe, a majority say *yes* (Nichols and Knobe 2007). It remains extremely controversial how we ought to interpret these findings. Nichols and Knobe suggest that the emotions aroused by contemplating morally bad actions cause

people to make a mistake: The emotion interferes with the application of their incompatibilist commitments. Murray and Nahmias (forthcoming) by contrast, think that the incompatibilist responses are actually the result of a confusion between *determinism* – which is a claim about causation – and *bypassing* of psychological capacities. They suggest that if we do not take care to avoid it, people will assume that a description of determinism entails that people's capacities for deliberation and choice are bypassed. That is, people misread the description of determinism as entailing that people will act as they are determined; *however, they decide to act* (rather than as entailing that people are determined to decide in one way rather than another). If we ensure that this confusion is avoided, Murray and Nahmias claim we reveal that most ordinary people are, after all, compatibilists.

The experimental philosophy movement is highly controversial in philosophy: The standard response from more traditional philosophers is to argue that what ordinary people think on first exposure to philosophical issues just is not relevant to what we ought to think once we have thought through the issues carefully. In his contribution to the volume, Tamler Sommers articulates a different worry: that the experimental philosophers are not probing the right intuitions. They are asking whether compatibilism seems right or not, when what they should be doing is asking whether ordinary people accept common (but controversial) *arguments* for compatibilism and incompatibilism. Sommers' point is that the philosophers involved in the free will debate never argue as follows:

1. It seems to me that if determinism is true, we lack free will.
2. Therefore, if determinism is true, we lack free will.

Rather, they present arguments for the incompatibility of free will and determinism. For instance, they argue:

1. If determinism is true, then what we do is the inevitable result of the laws of nature and the physical facts that pertain before we are born.
2. But we lack any kind of power over the laws of nature and the physical facts that pertain before we are born.
3. Therefore, we lack any power over what we do.

(Of course, compatibilists have arguments of their own, and rebuttals to incompatibilist's arguments). Sommers thinks we ought to see how ordinary people respond to these arguments, not how they respond to the simplistic "argument" that preceded it.

In response to Sommers, we might suggest that the more sophisticated arguments, like the one I have just given an extremely simplified version of, are attempts by incompatibilists to systematize and explain their pretheoretical intuition, which really is along the lines of premise 1 of the simplistic argument. If that is right, some case can be constructed for thinking that the intuitions of ordinary people with regard to the simplistic arguments are *also* worth probing. Sommers is certainly right in suggesting that that is not *all* experimental philosophers of free will should be doing. As he concedes, they are beginning to probe intuitions with regard to arguments, as well as continuing the kind of work he regards as fruitless.

Conclusion

The issues canvassed in the contributions to this section, and surveyed in this introduction, remain highly controversial. They are also extremely exciting. They perfectly illustrate the scope and the importance of neuroethics: They are issues that are directly relevant to important human concerns (are we responsible for our actions? Is criminal punishment justified? Ought we to feel pride and shame for what we do and what we are?) and require for their proper assessment both neuroscientific expertise and philosophical sophistication. The papers here represent current thinking by some of the most important contributors to their assessment. They should serve as a stimulus to further research on this significant neuroethical topic.

Cross-References

► Neurolaw: Introduction
► The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy

References

Caruso, G. (2012). *Free will and consciousness: A determinist account of the illusion of free will*. Lanham: Lexington Books.
Levy, N. (2011). *Hard luck*. Oxford: Oxford University Press.
Levy, N. (2014). *Consciousness and moral responsibility*. Oxford: Oxford University Press.
Libet, B., Gleason, C., Wright, E., & Pearl, D. (1983). Time of unconscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain, 106*, 623–642.
Mele, A. (2009). *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.
Murray, D., & Nahmias, E. (forthcoming). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*. DOI: 10.1111/j.1933-1592.2012.00609.x
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*, 663–685.
Soon, C. S., Brass, M., Henize, H.-J., & Haynes, J. D. (2008). Unconscious determinates of free decisions in the human brain. *Nature Neuroscience, 11*, 543–545.
Stillman, T. F., Baumeister, R. F., & Mele, A. R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology, 24*, 381–394.
Suarez, A., & Adams, P. (Eds.). (2013). *Is science compatible with free will? Exploring free will and consciousness in the light of quantum physics and neuroscience*. New York: Springer.
Tse, P. U. (2013). *The neural basis of free will: Criterial causation*. Cambridge, MA: The MIT Press.
Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*, 606–607.

Consciousness and Agency

14

Tim Bayne and Elisabeth Pacherie

Contents

Abstract

There are three projects within the cognitive science of agency and consciousness that are of particular interest to neuroethics: the *descriptive project*, the *genetic project*, and the *substantive project*. The descriptive project is concerned with characterizing our everyday experience of, and beliefs about, agency. What is the folk view of agency? The aim of the genetic project is to give an account of the psychological mechanisms involved in constructing our experience of, and beliefs about, agency. How is the folk view of agency to be explained? The substantive project is concerned with determining the degree to which our experiences of, and beliefs about, agency are correct and to what degree they

T. Bayne (✉)
Philosophy, School of Social Sciences, University of Manchester, Manchester, UK
e-mail: tim.bayne@manchester.ac.uk; tim.bayne@gmail.com

E. Pacherie
Institut Jean Nicod – UMR 8129, ENS, EHESS, CNRS, Paris, France
e-mail: pacherie@ens.fr

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_117,

might need to be revised in light of findings from the cognitive sciences. Is the folk view of agency basically correct or does it need to be modified in fundamental ways (as "will skeptics" argue)? This entry provides an overview of recent research relating to all three projects.

Introduction: Three Projects

It is possible to distinguish three distinct, but related, projects in the burgeoning literature on the relationship between agency and consciousness: the *descriptive* project, the *genetic* project, and the *substantive* project. The descriptive project is concerned with charting the contours of what, following Sellars (1962), can be described as the manifest image of human agency. How do we experience our own agency? What is the content of our intuitive, prescientific conception of human agency? The aim of the genetic project is to give an account of the psychological mechanisms involved in the construction of the manifest image of human agency. Why do we have the pre-theoretical conception of agency that we do? The focus of the substantive project is that of determining the degree to which the manifest image of agency is correct. To what extent might the folk conception of agency need to be revised in light of what we have learnt from the cognitive sciences?

Although these projects are formally distinct, there are many important connections between them. Most obviously, tackling the genetic project presupposes that one has made a certain degree of progress with the descriptive project, for until one has an account (if only partial) of the content of the manifest image of agency one won't be in a position to know quite what it is that the genetic project is meant to explain. There are also important points of contact between the genetic project and the substantive project, as we shall see in due course.

The Descriptive Project

The manifest image of agency comprises two kinds of states: agentive *experiences* and agentive *beliefs* (Bayne and Pacherie 2007). Agentive experiences involve the awareness of agency as such. Although some agentive experiences concern the agency of others, we will focus here – as the literature does – on agentive experiences that concern one's own agency (Horgan et al. 2003; Bayne 2008; Pacherie 2008; Wakefield and Dreyfus 1991). The literature contains a great variety of terminology here, with authors referring to experiences of deliberation, experiences of decision-making, experiences of intentionality, experiences of freedom, experiences of mental causation, the awareness of movement, the awareness of intentions to act, the sense of control, the sense of effort, and the sense of agency (among many others).

It is possible to impose some much-needed structure on this discussion by distinguishing two kinds of elements associated with agentive experience: those that are associated with *all* agentive experiences ("core elements") and those

that are associated with only *some* agentive experiences ("non-core elements") (Bayne 2010). One core element of agentive awareness is the sense of oneself as acting. Arguably, this experience cannot exist independently of any experience of what it is one is doing – as a feeling simply "floating in the air" so to speak – but must be accompanied by a sense of *what* it is one is doing, whether that be moving one's body in a certain way (e.g., moving a finger), performing an action of particular kind (e.g., pressing a key), or trying to realize a certain goal (e.g., starting a new paragraph). Having an agentive experience does not require one to identify the action the experience is about at a particular level of specification, but it does demand that the action be specified in some way, however vague. And if that is right, then agentive experience includes two core elements: a sense of agency (i.e., an awareness of oneself as acting) and some specification of the action the sense of agency is directed towards.

What about the non-core elements of agentive experience? Among the various kinds of experiential states associated with agency are experiences of effort, experiences of deliberation, experiences of decision-making, and experiences of freedom. These experiences do not appear to characterize all forms of agency – for example, spontaneously and unreflectively cleaning one's glasses or brushing the hair from one's eyes might not be accompanied by any sense of effort, deliberation, decision-making, or freedom – but they do seem to accompany some of our actions. However, giving a precise characterization of the content of these agentive experiences has proven to be very challenging. For example, although many theorists would agree that there is a sense in which we experience ourselves as free, there is little agreement as to what exactly this sense of freedom involves (Nahmias et al. 2004).

Let us turn now to the other component of the manifest image of agency: agentive belief. We take this label to include all our pre-theoretical (or "folk") beliefs about agency, whether they are held explicitly or whether they are merely implicit in our intuitive responses to situations. Examples of such beliefs might include the belief that we possess free will, that consciousness plays an important role in the control of intentional actions, that deliberation typically improves the quality of decision-making, and so on. However, these characterizations of the folk conception of agency are fairly rough and ready, and it would be nice to have a more precise analysis of what the folk conception of agency involves. A great deal of recent work that has been conducted under the "experimental philosophy" label has attempted to meet this need (see Sommers 2010 for a review). Two issues in particular have been the focus of recent discussion: (1) What is the folk conception of free will and (2) What is the folk conception of intentional agency?

One of the central debates in the philosophy of action is that between compatibilists and incompatibilists. Compatibilists hold that free will and determinism are compatible with each other, whereas incompatibilists hold that they are incompatible with each other. The question that concerns us here is not how free will *should* be analyzed (whatever precisely that might mean) but what conception of free will the folk have. Are the folk compatibilists or incompatibilists?

Attempts to answer this question paint a rather mixed picture (Nichols 2011). Eddy Nahmias and his colleagues have argued that the folk have a predominantly

compatibilist conception of free will (Nahmias et al. 2005; Turner and Nahmias 2006). They presented undergraduates with vignettes outlining a world in which human agency (along with everything else) was perfectly predictable and asked them whether free will was possible in such a world. They found that 76 % of their subjects claimed that free will was possible in such a world, suggesting that the folk are "natural compatibilists."

However, other studies have suggested that the folk have a predominantly incompatibilist conception of free will. In one study, Nichols (2004) presented a group of 4- and 5-year-olds with scenarios involving physical events and moral choices and asked them questions on the model of the following:

Moral choice: "If everything in the world was the same right until she chose to steal, did Mary have to choose to steal?"

Physical event: "If everything in the world was the same right until the time that the water boiled, did the water have to boil?"

Nichols found that the children were more likely to say that the physical events had to happen than they were to say that the moral choice events had to happen, and he concluded that the folk are intuitive incompatibilists.

In work designed to examine the tension between these findings, Nichols and Knobe (2007) explored the idea that folk responses to questions about free will depend on precisely how such questions are framed. Nichols and Knobe described two universes to their subjects, a fully deterministic universe (A) and a universe in which everything except human decisions are deterministically fixed (B). The subjects are first asked which of these two universes most closely resembles our own, with over 90 % of subjects choosing universe B. Subjects are then split into two conditions: an abstract condition and a concrete condition. In the former, subjects are asked whether it is possible for someone in universe A to be fully morally responsible for their actions, with 86 % of the subjects saying "no." But in the concrete condition, 72 % of subjects said that a man in universe A who kills his wife and three children in order to be with his secretary is morally responsible for his actions. In other words, the concreteness (and in particular affective character) of a scenario has an impact on the judgments that subjects make about it.

Just what we should conclude from all of this is, as yet, quite unclear. One possibility is that there is no standard folk conception of free will; instead, significant numbers of folks are compatibilists and significant numbers are incompatibilists; indeed, there is some evidence that individual differences in conceptions of free will covary with personality differences (Feltz and Cokely 2009). Another possibility is that (most) of us don't have a single unified conception of free will but instead have a view of free will that incorporates both compatibilist and incompatibilist elements. Deciding between these (and other) possibilities will no doubt be on the agenda of experimental philosophy for some time to come.

Let us turn now to the folk conception of intentional action. A central question in the analysis of intentional agency is whether foreseen side effects of an action are ever brought about intentionally. Some philosophers have argued that side effects are never brought about intentionally, whereas others have argued that side effects can be brought about intentionally. In an influential series of studies,

Joshua Knobe (2003a, b; see also Malle and Knobe 1997) asked what the folk conception of intentional action is. Knobe discovered that people are much more likely to ascribe intentionality to side effects when those effects are negative (e.g., harming the environment) than when they are positive (e.g., helping the environment). Interestingly, the folk conception of intentional agency is very different from the one which philosophical treatments of intentional agency would have led one to predict.

The Genetic Project

The second of the three projects that we have distinguished is the genetic project: what accounts for the content of the manifest image of agency?

One important question here concerns the relationship between agentive experience and agentive belief. It is plausible to suppose that the content of agentive experience plays a central role in determining the content of agentive belief and that much of what we intuitively believe about agency is grounded in our first-person experiences of agency. On this view, explaining why the manifest image of agency has the content that it does is primarily a matter of explaining why we have the agentive experiences that we do.

This "empiricist" account of the folk conception of agency is not the only option on the table. Another possibility is that certain folk beliefs about agency might be grounded in our moral commitments. For example, Nichols (2004) has suggested that the incompatibilist elements in the folk view of free will might derive from perceived constraints on moral responsibility. The idea is that we reason as follows: we are subject to moral requirements; we could be subject to moral requirements only if we possessed incompatibilist free will; so we possess incompatibilist free will. Note that Nichols does not endorse this argument but merely suggests that it might play a role in explaining the widespread appeal of incompatibilism.

Let us leave these questions to one side and turn to the question of how agentive experience itself might be generated, for it is this issue that has received the lion's share of attention in the recent literature. A number of accounts of agentive experience can be distinguished, but they all have a shared commitment to the idea that the sense of agency is produced when there is a match between cues x and y. What distinguishes these accounts from each other is their conception of (1) the nature of the cues being compared; (2) the nature of the processes involved in the production of the sense of agency; and (3) how closely these processes are related to action production and control processes.

Two positions define the two ends of the spectrum of possibilities: the motor prediction view and the cognitive reconstruction view. On the motor prediction view, the sense of agency is generated by processes dedicated to action control. On the cognitive reconstruction view, the sense of agency is generated by a general-purpose process of retrospective causal inference.

The motor prediction view is inspired by computational theories of motor control. According to these theories, when the motor system generates a motor

command, an efference copy of this command is sent to forward models whose role is to generate predictions about its sensory consequences in advance of actual execution. Error signals arising from the comparison of desired, predicted, and actual states (as estimated from sensory reafferences) are used to make corrections and adjustments. The motor prediction view holds that the signals used for motor control also provide cues to agency (Frith et al. 2000). In particular, it holds (1) that awareness of initiating an action is based on a representation of the predicted consequences of making that action, rather than its actual consequences, and on the congruence between the predicted state and the desired state and (2) that for this experience of agency to continue, the predicted consequences must remain congruent with the sensory reafferences when they become available.

Claim (1) – and therefore the possibility that the sense of agency can emerge in advance of actual sensory effect and be based on premotor processes alone – is supported by evidence that awareness of initiating a movement in healthy subjects occurs between 80 and 200 ms before the movement actually occurs (Libet et al. 1983; Libet 1985; Haggard and Eimer 1999). Evidence for claim (2) – that the sense of agency also depends on the congruence between predictions and sensory reafferences – comes from studies where these reafferences are artificially manipulated by introducing temporal delays and spatial distortions of feedback. These studies demonstrate that the sense of agency is gradually reduced as these discrepancies increase (Fourneret and Jeannerod 1998; Knoblich and Kircher 2004; Sato and Yasuda 2005).

In contrast, the cognitive reconstruction view downplays the contribution of the motor system to the sense of agency and proposes that it is inferred retrospectively from the existence of a match between a prior thought and an observed action. Thus, on Wegner's "theory of apparent mental causation" (Wegner 2002), a general-purpose causal inference process is at play. If an action is consistent with a prior thought of the agent and other potential causes of the action are not present or salient, a sense of agency for the action will be induced.

There is empirical evidence that high-level inferential processes play a role in determining the sense of agency for an action. Studies of Wegner and colleagues have demonstrated that cognitive cues can alter the sense of agency for an action independently of changes in sensorimotor and perceptual cues. For instance, in their "I-Spy" study (Wegner and Wheatley 1999), a participant and a confederate of the experimenter had joint control of a computer mouse that could be moved over any one of a number of pictures on a screen. When participants had been primed with the name of an item on which the mouse landed, they expressed a stronger sense of agency for the action of stopping on that object (when in fact the stop had been forced by the confederate). Further studies also suggest that subliminally priming an outcome just before the outcome is produced can enhance the sense of agency for that outcome (Aarts et al. 2005) and that priming an outcome relatively far in advance of it can augment the sense of agency, but only if the outcome is attached to positive affect (Aarts et al. 2009).

There is now a growing consensus that the motor prediction view and the cognitive reconstruction view are not mutually exclusive but complementary and that intrinsic cues (cues provided by the motor system) and extrinsic cues (such as cognitive primes) both contribute to the sense of agency (Pacherie 2008; Sato 2009; Synofzik et al. 2008; Moore et al. 2009; Moore and Fletcher 2012). Researchers are now trying to develop integrative frameworks in order to get a better understanding of how all these agency cues interact.

One way to try to combine the motor prediction view with the cognitive reconstruction view is to appeal to the distinction between pre-reflective agentive experiences and reflective agentive beliefs or judgments (Bayne and Pacherie 2007; Gallagher 2007; Haggard and Tsakiris 2009) and to argue that while motor processes contribute mainly to feelings of agency, interpretive processes contribute mainly to judgments of agency. This conceptual distinction is echoed methodologically in the ways agency is measured in experimental studies. While some studies (Farrer et al. 2003; Metcalfe and Greene 2007; Sato and Yasuda 2005) investigate agency by asking participants to explicitly judge whether they caused a particular sensory event, other studies use implicit agency measures such as intentional binding and sensory suppression. Intentional binding is a phenomenon, first reported by Haggard and his colleagues (2002), whereby an action and its external sensory consequences are compressed together in subjective time. As intentional binding occurs only for voluntary actions (Tsakiris and Haggard 2003) and is furthermore modulated by the statistical relation between events (Moore and Haggard 2008), it is considered to provide an implicit measure of agency. Sensory attenuation of self-produced action effects has also been used as an implicit measure of agency. When the internally generated motor predictions about the sensory consequences of one's ongoing actions and their actual sensory consequences are congruent, the sensory percept is attenuated, thereby enabling a differentiation between self-generated and externally generated sensory events (Blakemore et al. 2002; Cullen 2004). However, recent studies showing that prior authorship beliefs can modulate both sensory attenuation and intentional binding (Desantis et al. 2011, 2012) suggest that drawing a sharp distinction between feelings of agency supported by motor processes and judgments of agency supported by interpretive processes may be oversimplistic.

A promising approach is to appeal to a Bayesian integrative framework involving a hierarchy of prediction and model building. Thus, Moore and Fletcher (2012) propose that the sense of agency is determined by a Bayesian process of cue integration, where the predictions generated at higher levels of the hierarchy provide the priors for the lower levels, i.e., constrain the interpretation of cues available at lower levels. In this model, cue integration is itself the product of both the strength of the priors and the weights attached to the available cues as a function of their reliability. When priors are weak – as, for example, when one is quite unsure what the effects of pressing a button will be – one may still have a strong sense of agency for the ensuing consequence, provided that perceptual reafferences carrying information about it are very reliable. Conversely, if my

priors are very robust, I may have a strong sense that I produced a certain effect in the world, even though the feedback I get is weak or ambiguous. When both priors and reafferent cues are weak, my sense of agency may be correspondingly weakened. While this Bayesian approach does not allow for a sharp distinction between agentive experiences and agentive judgments, it can accommodate the idea that high-level priors exert more influence on agentive judgments than on agentive experiences.

The Substantive Project

We turn now to the third of three projects that we outlined in the introduction: the substantive project. Briefly put, the aim of the substantive project is to determine the degree to which the manifest image of agency is correct. How accurate are our experiences of and beliefs about agency? The substantive project has dominated discussions of the relationship between agency and consciousness over the last two or so decades. At the center of these discussions is a position that has been dubbed *will skepticism*. Will skeptics argue that important elements of the manifest image of agency are at odds with the scientific image of agency and as such should be rejected or at least revised.

The most popular form of argument for will skepticism attempts to put pressure on the folk conception of agency by trying to show that some of its commitments are false. For example, arguments for will skepticism that appeal to Libet's (1985) influential studies of the readiness potential claim that these studies are at odds with the folk commitment to the idea that freely willed actions are initiated by an act of conscious volition.

These arguments share the following "two-pronged" structure. The first prong involves the claim that the folk conception of agency is committed to ascribing a certain feature ("feature X") to human agency. The second prong involves a claim to the effect that human agency does not in fact possess feature X. Cognitive science is relevant to the evaluation of each of these two prongs. Most obviously, it is relevant to the evaluation of the second prong, for the claim that human action lacks certain features is subject to the tribunal of empirical inquiry. But cognitive science is also relevant to the evaluation of the first prong, for the question of what precisely the folk conception of agency is committed to is a matter of the descriptive project, and that – as we have seen – falls within the domain of cognitive science.

The argument for will skepticism that appeals to Libet's experiments regarding free will and the readiness potential (Libet 1985; Libet et al. 1983) is one of the most the most widely discussed in the current literature (see, e.g., Banks and Pockett 2007; Mele 2009; Nahmias 2013; Sinnott-Armstrong and Nadel 2010). In these experiments, subjects were asked to flex their wrist at will and to note when they felt the urge to move by observing the position of a rapidly rotating dot on a special clock. While subjects were both acting and monitoring their urges (intentions, decisions) to act, Libet used an EEG to record the activity of

prefrontal motor areas. On average, participants reported the conscious intention to act, what Libet called the W judgment, about 200 ms before the onset of muscle activity. By contrast, the EEG revealed that preparatory brain activity, termed by Libet the type II readiness potential (RP), preceded action onset by about 550 ms. In other words, their brain started preparing the action at least 350 ms before the participants became aware of the intention to act. In fact, for reasons that we need not explore here, Libet claimed that this gap was likely to be closer to 400 ms in length.

As a number of commentators have pointed out, Libet's paradigm is subject to a number of methodological problems (see, e.g., the commentaries on Libet 1985). To take just one example of these problems, Libet's paradigm requires subjects to divide their attention between the position of dot on the clockface and their own agency. The demand to divide one's attention between two perceptual streams in this way is a notorious source of error in temporal-order judgments. Despite these difficulties, Libet's basic findings have been replicated by a number of laboratories using studies that are free of these methodological difficulties.

Although there is some variability between studies, the claim that "Libet actions" – that is, simple and (relatively) spontaneous motor actions – involve an RP whose onset precedes the time of the subjects' W judgment by about 400 ms or so is largely undisputed. What is in dispute are the implications of these results for questions concerning free will.

Libet denied that his results establish free will skepticism, for he argued that the gap of 150 ms between the agent's conscious decision and the onset of the action allowed for a kind of free will in the form of conscious veto. However, many theorists have seen in Libet's work the death knell of free will. In their review of his work, Banks and Pockett (2007, p. 658) describe Libet's experiments as providing "the first direct neurophysiological evidence in support of [the idea that perceived freedom of action is an illusion]."

Unfortunately, few skeptics have said exactly how Libet's data are supposed to undermine free will. Here is one way in which Libet's data might be thought to put pressure on free will (Bayne 2011):

1. The actions studied in the Libet paradigm are not initiated by conscious decisions but are instead initiated by the RP.
2. In order to exemplify free will, an action must be initiated by a conscious decision.
3. So, the actions studied in the Libet paradigm are not freely willed. (From [1] and [2]).
4. Actions studied in the Libet paradigm are central exemplars of free will (as intuitively understood), and so if these actions are not freely willed, then no (or at least very few) actions are freely willed.
5. So no human actions are freely willed. (From [3] and [4]).

We will refer to this as *the skeptical argument*. The skeptical argument is valid, so if it is to be resisted, we need to reject one (or more) of its premises. Let us begin by considering (4).

The Scope of Free Will

Are the actions that form the focus of the skeptical argument – “Libet-actions” – paradigm examples of our intuitive notion of free will? Libet himself had no doubts about the answer to this question – he took himself to have studied an “incontrovertible and ideal example of a fully endogenous and ‘freely voluntary’ act” (Libet et al. 1983, p. 640) – but not everyone shares this view. Adina Roskies, for example, claims that Libet-actions are at best “degenerate” examples of free will and suggests that we ought to focus on actions that are grounded in our reasons and motivations if we are interested in “how awareness and action are related insofar as they bear on freedom and responsibility” (2010b, p. 19).

To make progress here we need a taxonomy of action types. One useful distinction is between *automatic* actions and *willed* actions. Automatic actions flow directly from the agent’s standing intentions and prepotent action routines. Many of our everyday actions – washing the dishes, answering the telephone, and reaching for a door handle – are automatic. Our awareness of various features of our environment together with overlearned action schemas conspires to trigger the appropriate actions with only the minimal participation of conscious deliberation or decision on the part of the agent. Willed actions, by contrast, require the intervention of executive processes. Some willed actions – what we call “deliberative actions” – involve only decision. Consider the experience of finding oneself in a restaurant confronted by a number of equally attractive – or, as the case may be, unattractive – options on the menu. One needs to make a choice, but it does not matter what one orders. Other willed actions – what we call “deliberation actions” – involve both decision and deliberation. Consider Sartre’s case of the young man who must choose whether to look after his aged mother or join the resistance. Here, the function of decision-making is not to select from amongst a range of options between which one is relatively indifferent (as is the case in contexts of disinterested actions) but to draw on one’s reasons in making a good decision.

Are Libet-actions automatic or willed? Although they are embedded in a wider agentive context – a context that includes a conscious decision to produce an action of a certain type within a certain temporal window – Libet-actions are not plausibly regarded as automatic. Unlike standard examples of automatic actions, Libet actions are not triggered by an external cue. They may not be the “ideal examples” of fully spontaneous agency that Libet took them to be, but Libet-actions do seem to be genuine instances of willed agency nonetheless.

But although Libet-actions involve an act of will, they do not involve deliberation – at least, not immediately prior to the action. They are “disinterested” rather than “deliberative” actions, for the agent has no reason to flex their wrist at one particular time rather than another or to flex it in one way rather than another. Indeed, Libet experiments are explicitly constructed so as to minimize the rational constraints under which the subject acts. We might think of Libet-actions as manifesting the liberty of indifference.

With the foregoing in hand, let us return to the question of whether Libet-actions are paradigms of free will (as we intuitively conceive of it). Are disinterested actions our central exemplars of free will, or does that epithet belong to deliberative actions? Philosophers do not agree on the answer to this question, and the systemic research that would be required in order to settle this dispute has not been carried out. That being said, we suspect that Roskies is right to identify the central or core cases of free will – at least, the kind of free will that is most intimately related to moral agency – with deliberation and rational reflection.

But even though Libet-actions might not be paradigms of free agency, it seems clear that they *do* fall within the scope of our pre-theoretical notion of free will. As such, the free will skeptic is perfectly within his or her rights to claim that if Libet actions – and indeed disinterested actions more generally – are not free, then an important component of our commonsense conception of free will would be threatened. In sum, although (4) is unacceptable as stated, the skeptical argument is not thereby rendered impotent, for the question of whether Libet-actions manifest free will is itself an important one. Libet actions might not qualify as ideal examples of free will, but they do provide the free will skeptic with a legitimate target.

The Initiation of Free Action

Let us turn now to the second premise of the skeptical argument:

> (2) In order to exemplify free will, an action must be initiated by a conscious decision.

We can think of (2) as the "conceptual" step of the skeptical argument, for its plausibility turns chiefly on the contours of our everyday (or "folk") notion of free will.

But is (2) true? In order to engage with this question, we need to consider what it means for an action to be initiated by a conscious decision. According to one view, an action is initiated by a conscious decision only if it has its point of origin in a conscious decision that is itself uncaused. Is this how we should understand (2)?

Certain incompatibilists might think so. More to the point, certain kinds of incompatibilists might argue that the *folk* are implicitly committed to this claim and thus any evidence to suggest that this claim is false would require that our folk conception of free will be modified in some way. However, as we noted in discussing the descriptive project, it is far from clear just what "the folk" conception of free will is. Although the folk can be prompted to give incompatibilists responses in certain contexts, they can also be prompted to give compatibilists responses in others, and thus it remains an open question just how deeply committed the folk are to incompatibilism.

Let us turn to another reading of (2). One might argue that all it is for an action to be initiated by a conscious decision is for that action to have its point of origin in that decision, without also requiring that that decision is itself an uncaused event. Is (2) plausible even on this weaker reading of it?

Note first that the very idea that an action can always be traced back to a *single* point of origin is open to challenge. Rather than thinking of actions as originating with particular discrete events, we might do better to conceive of them as the outcome of multiple events and standing states, no single one of which qualifies as "the" point of origin of the action. Just as the Nile has more than one tributary, so too many of our actions might result from multiple sources.

Secondly, to the extent that free actions can be traced back to a point of origin, it is by no means obvious that this point of origin must always be a conscious decision (Levy 2005). Consider a thoughtless comment that is uttered on the spur of the moment and without forethought. Despite the fact that such an utterance is not consciously initiated, one might think that the notion of free will has some kind of grip in such contexts. But, the objection continues, if that is right, then (2) is too demanding: freely willed actions need not be initiated by conscious decisions.

In response to these points, the advocate of the Libet argument might argue that even if it's not an essential feature of all freely willed actions that they have their point of origin in a conscious decision, it is a feature of the kinds of (supposedly free) actions that *Libet* studied. Unlike those actions that we experience as automatic, Libet-actions are accompanied by the "phenomenology of conscious initiation": one experiences oneself as deciding to act here-and-now. And, the will skeptic might continue, if the neural data demonstrate that the action has been initiated before the agent is aware of their decision, then this sense of origination is illusory.

Some authors would take issue with this characterization of the agentive experience that accompanies Libet-actions. For example, Terry Horgan (2010) acknowledges that one would experience oneself as beginning to actively undertake an action at some specific moment in time, but he denies that this phenomenology would involve any sense that one's behavior is caused by one's mental states. Instead, he suggests, one would experience oneself as "authoring" the behavior. Horgan's comments raise deep and important issues, but we lack the space to engage with them here. Instead, we will turn to the first premise of the Libet argument and the question of whether Libet-actions really are initiated by the readiness potential rather than the agent's conscious decision.

Conscious Decisions and the Readiness Potential

The first premise of the skeptical argument is as follows:

> (1) The actions studied in the Libet paradigm are not initiated by conscious decisions but are instead initiated by the RP.

In order to evaluate (1), we need to consider what it is for an event to initiate an action. Let us say that ε initiates α only if there is a robust correlation between ε-type events and α-type events, such that in normal contexts there is a high probability that an ε-type event will be followed by an α-type event. (The notion of origination clearly requires more than this, but it is implausible to suppose that it requires less than this.) So, if the RP initiates the agent's action, then we ought to expect RP events to be "immediately" followed by the appropriate action, unless something unusual happens (such as the person being struck by lightning). Or, to put it the other way around, we should expect that when there is no movement, there is also no RP event. Is this the case?

As several commentators have observed, the back-averaging techniques used to measure RPs do not allow us to answer this question. Because the RP on any one trial is obscured by neural noise, what is presented as "the RP data" is determined by averaging the data collected on a large number of trials. In order to compute this average, the EEG recordings on different trials need to be aligned, and this requires some fixed point – such as the onset of muscle activity or some other observable behavior on the part of the subject – that can be identified across trials. This technique has two main drawbacks. First, as Roskies (2010b) and Trevena and Miller (2002) note, because it involves averaging across a number of trials, certain aspects of the data might be statistical illusions. In other words, features of the relationship between (say) the RP and the W judgment might characterize the averaged data even thought they do not characterize any of the individual trials that contribute to that grouped data. Second, because action onset serves as the needed fixed point for the alignment of EEG recording, any RPs that are not followed by an action simply won't be measured, and so we don't know how robust the correlation between the RP and Libet actions is (Mele 2009).

There are indirect reasons for thinking that the relation between the RP and subsequent action may not be as tight enough for the RP to qualify as the point of origin of the action. Firstly, we know that the nature of the experimental context can significantly affect both the temporal properties and the strength of the RP signal. Subjects who are highly motivated to perform the task produce a large RP, whereas the RP almost disappears in subjects who have lost interest in the task (McCallum 1988; Deecke et al. 1973; see also Rigoni et al. 2011). Secondly, it is possible to make willed responses to stimuli in very much less than 550 ms, which indicates that a type II RP is not "the" point of origin even where it occurs. Thirdly, another neural event – the *lateralized* readiness potential (LRP) – appears to be more strongly coupled to agency than the (generalized) RP is. Whereas the (generalized) RP is symmetrically distributed over both hemispheres, the LRP is restricted to the hemisphere contralateral to the hand that is moved. Haggard and Eimer (1999) found that the LRP was more robustly correlated with the subsequent action than the RP as well as tightly coupled to the W judgments that subjects make. However, a version of the Libet argument in which (1) is replaced with a corresponding claim about the LRP does not possess even the surface plausibility that (1) does. (Note, however, that a recent study by Schlegel et al. (2013) failed to replicate Haggard

and Eimer's finding, and found no within-subject covariation between LRP onset and W judgment, leading them to conclude that neither RP onset nor LRP onset cause W).

In a recent experiment, Schurger and colleagues (2012) used a modified Libet task to circumvent the limitations due to back-averaging techniques. Their aim was to test the proposal that RPs correlate with predecision activity rather than with activity which coincides with, or is subsequent to, the agent's decision (as Libet thought). Schurger and colleagues proceeded by assuming that the decisions of the participants in Libet's experiment can be modelled – as neural decision tasks typically are – in terms of an accumulator-plus-threshold mechanism: decisions are made when relevant evidence accumulated over time reaches a threshold. What is unique to Libet's task is that subjects are explicitly instructed not to base their decision on any specific evidence. Schurger and colleagues propose that the motor system constantly undergoes random fluctuations of RPs and that this random premotor activity is used as a substitute for actual evidence. According to their stochastic decision model, the decision process, given Libet's instructions, amounts to simply shifting premotor activation up closer to the threshold for initiation of the movement and waiting for a random threshold-crossing fluctuation in RP. Time-locking to movement onset ensures that these fluctuations appear in the average as a gradual increase of neuronal activity, when in fact what is measured are simply random fluctuations of RPs that happened to cross a decision threshold.

Thus the two models predict the same premotor activation buildup when a movement is produced, but whereas on Libet's postdecision interpretation of this buildup there should be no premotor activity (and hence no RPs) when no movement is produced, on the predecision interpretation there should be continuous random fluctuations in RPs even when no movement is produced. Schurger and colleagues reasoned that it should be possible to capture these fluctuations by interrupting subjects in a Libet task with a compulsory response cue and sorting trials by their reaction times. On the assumption that the interrupted responses arise from the same decision accumulator as the self-initiated ones, response times should be shorter in trials in which the spontaneous fluctuations of RPs happened to be already close to threshold at the time of the interruption. On the assumption that close to threshold activity reflects spontaneous fluctuations of RPs rather than mounting preparation to move building over the course of the entire trial, slow and fast reaction times should be distributed equally across time within trials. To test these predictions, they devised what they called a *Libetus interruptus* task, where they added random interruptions to trials. They found, as they had predicted, that slow and fast responses to interruptions were distributed equally throughout the time span of the trial.

According to the predecision model, Libet's contention that the neural decision to move happens much before we are aware of an intention or urge to move is unfounded. The neural decision to move isn't made when a RP starts building up, since spontaneous fluctuations of RPs happen all the time, but when a random fluctuation in RP crosses a threshold. The reason we do not experience the urge to move earlier is simply that the decision threshold has not yet been crossed and thus the decision has not yet been made. While Schurger and colleagues take no stand on

the exact temporal relation between the conscious urge to move and the neural decision to move, their results cast serious doubt on Libet's claim that the neural decision to move coincides with the onset of the RP and thus on his further claim that since RP onset precedes the urge to move by 350 ms or more, conscious intentions play no role in the initiation of the movement. If instead the neural decision to move coincides with a much later threshold-crossing event, it remains at least an open possibility that this event coincides with and constitutes the neural basis of a conscious urge to move. In any case, Schurger and colleagues also insist that this threshold-crossing event should not be interpreted as *the* cause of the movement but rather as simply one of the many factors involved in the causation of self-initiated movements.

Taken together, these points suggest that the RP is unlikely to qualify as "the" point of origin of the action. If the RP has a psychological interpretation – and it is far from clear that it does – then we should perhaps think of it as the neural correlate of an "urge" or "inclination" to act, rather than as the neural basis of the decision to act now (Gomes 1999; Mele 2009). The RP may be one of the many tributaries that contribute to an action, but it is not its "origin" in any intuitive sense of that term.

Other Sources of Will Skepticism

In a series of papers and most influentially in his book *The Illusion of Conscious Will*, Daniel Wegner has argued that central components of the folk psychological conception of agency are inaccurate and should be jettisoned. As he puts it, the conscious will "is an illusion." Precisely what Wegner means by describing the conscious will as an illusion is open to some debate (Bayne 2006; Mele 2009; Nahmias 2002), but we take his central claim to be this: agentive experience misrepresents the causal path by means of which one's own actions are generated.

One reason that Wegner gives for thinking that the conscious will "is an illusion" involves the idea that agentive experiences are theoretically mediated. As he puts it, "[Conscious will is an illusion] in the sense that the experience of consciously willing an action is not the direct indication that the conscious thought has caused the action" (Wegner 2002, p. 2). As we have seen, there is very good reason to think that agentive experiences are theoretically mediated, but it is difficult to see why will skepticism should follow from this. Even if the folk are intuitively inclined to think that our access to our own agency is direct and unmediated – and we are far from certain that such a view *is* part of the folk conception of agency – there is little reason to think that such a view is part of the core conception of agency.

Another sense in which one might regard the conscious will as illusory involves the idea that experiences of doing are systematically, or at least frequently, *non-veridical*: experiences of doing misrepresent our agency and the structure of our actions. This seems to be Wegner's main line of argument for will skepticism, and *The Illusion of Conscious Will* contains extensive discussion of cases that are alleged to involve dissociations between the exercise of agency and the phenomenology of agency. Some of these cases appear to demonstrate that we can experience ourselves

as doing something that someone else is doing (and that we are not). Wegner calls such cases *illusions of control*. The I-Spy experiment (discussed earlier) is an example of an illusion of control. Other dissociations involve experiencing someone (or something) else as the agent of what one is doing. Wegner calls such phenomena *illusions of action projection*. Among the most fascinating of the various illusions of action projection that he discusses is facilitated communication, a practice that was introduced as a technique for helping individuals with communication difficulties. Facilitators would rest their hands on the hands of their clients as the client typed a message. Although the facilitators experienced no sense of authorship towards the typed message, there is ample evidence that the content of "facilitated" messages derived from the facilitator rather than the client (Wegner et al. 2003).

How might these dissociations support the case for will skepticism? On one reading of his argument, Wegner is mounting an inductive generalization: since some experiences of conscious will are non-veridical, it is reasonable to infer that most, and perhaps even all, such experiences are. But this argument seems weak, for the fact that experiences of agency *can* be non-veridical shows that the mechanisms responsible for generating such experiences are *fallible* but it does not show that they are *unreliable*. Another way to read the argument from dissociations is as an inference to the best explanation. The argument proceeds as follows: Since the phenomenology of agency plays no direct role in the genesis of action where such experiences are absent, we have good reason to think that it plays no direct role in the genesis of action when such experiences are present. As Wegner himself puts it, "If conscious will is illusory, automatisms are somehow the 'real thing', fundamental mechanisms of mind that are left over once the illusion has been stripped away. Rather than conscious will being the rule and automatism the exception, the opposite may be true" (2002, p. 143).

What does it mean to say that automatisms are the *fundamental* mechanisms of mind? To the extent that automatisms are action-generation procedures that do not involve intentional states of any kind then there may be a tension between automaticity and the experience of conscious will, but Wegner provides little evidence for the view that our actions are usually automatic in this sense of the term. If, on the other hand, automatisms are action-generating procedures that are nonconsciously initiated, then there is ample reason to describe much of what we do as automatic in nature. But on this conception of an automatism there is no conflict between automaticity and the feeling of doing. So there is no argument from automaticity (thus conceived) to the claim that the experience of conscious will is an illusion.

We do not deny that the phenomenology of agency *can* be illusory. Consider, for example, the experience of intentionality. An experience of intentionality will be non-veridical if the action in question is not guided by an intention or if it is guided by an intention other than the one that it seems to have been produced by. The phenomenon of confabulation suggests that at least one if not both of these conditions occur. But I think that there is little reason to assume that either kind of mistake is at all common in everyday life. Confabulation is striking precisely because it is unusual.

In fact, Wegner's own account of the genesis of the experience of doing suggests that such experiences will normally be veridical. According to the matching model, we experience ourselves as doing X when we are aware of our intention to X as being immediately prior to our X-ing and when we are not aware of any rival causes of our X-ing. Now, if we experience ourselves as having an intention to X, then it probably is the case that we do have the intention to X. (After all, it seems reasonably to suppose that introspection is generally reliable. At any rate, Wegner is not in a position to challenge the reliability of introspection, for he himself assumes the reliability of introspection insofar as he takes subjects to be reliable in reporting their experiences of conscious will.) But if one has an intention to X, and if one has in fact X-ed, and if one's intention to X is immediately prior to one's X-ing, then it is highly likely that one's intention to X is involved in bringing about one's X-ing. It would be remarkable if one had an intention to raise one's hand just prior to raising one's hand, but the intention played no causal role in the raising of one's hand. Far from showing that experiences of agency are illusory, Wegner's own model of how such experiences are generated predicts that they will normally be veridical.

Conclusion

Discussions of human agency in the last decades have been dominated by the issue of whether the scientific image of human agency undermines the manifest image of human agency and in particular its commitment to free will. For the most part, cognitive scientists have argued that the folk view of agency is undermined by cognitive science, whereas philosophers have generally denied that there is any such tension, either on the grounds that the scientists in question have misinterpreted the neuroscientific findings or on the grounds that they have assumed a tendentious account of the folk conception of agency.

Although neither camp can claim a decisive victory in this debate, it has prompted theorists to take a closer and more nuanced look at both agentive belief and experience, showing them to be more subtle and less monolithic than previously thought. It is also clear that both elements of the manifest image of agency have complex etiologies and that the sense of agency in particular depends on the integration of multiple cues. Finally, recent research in the neuroscience of decision and action control reveals both the multilevel structure of these processes and the flexibility of their organization.

These advances suggest that we should resist the temptation to think in dichotomous terms and that we may need to replace such questions as whether we have free will, whether we are natural incompatibilists, and whether our sense of agency is veridical with questions that do not lend themselves to yes/no answers. As Roskies (2010a) notes, it is also clear that the time is ripe for philosophers, cognitive scientists, and neuroscientists to engage in more constructive endeavors and use all of the tools at their disposal to advance a positive conception of human agency, one that does justice both to its strengths and to its limitations.

Acknowledgments We gratefully acknowledges the support of the Institute of Advanced Studies at the Central European University in Budapest (Pacherie) and the John Templeton Foundation (Bayne).

Cross-References

▸ Free Will and Experimental Philosophy: An Intervention
▸ Free Will, Agency, and the Cultural, Reflexive Brain
▸ Mental Causation
▸ Neuroscience, Free Will, and Responsibility: The Current State of Play

References

Aarts, H., Custers, R., & Wegner, D. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition, 14*(3), 439–458.

Aarts, H., Custers, R., & Marien, H. (2009). Priming and authorship ascription: When nonconscious goals turn into conscious experiences of self-agency. *Journal of Personality and Social Psychology, 96*, 967–979.

Banks, B., & Pockett, S. (2007). Benjamin Libet's work on the neuroscience of free will. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness*. Malden: Blackwell.

Bayne, T. (2006). Phenomenology and the feeling of doing: Wegner on the conscious will. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior?* (pp. 169–186). Cambridge, MA: MIT Press.

Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass, 3*, 1–21.

Bayne, T. (2010). Agentive experiences as pushmi-pullyu representations. In J. Aguilar, A. Buckareff & K. Frankish (Eds.), *New Waves in the Philosophy of Action*. Palgrave Macmillan, pp. 219–36.

Bayne, T. (2011). Libet and the case for free will scepticism. In R. Swinburne (Ed.), *Free will and modern science* (pp. 25–46). Oxford: Oxford University Press.

Bayne, T., & Pacherie, E. (2007). Narrators and comparators: The architecture of agentive self-awareness. *Synthese, 159*, 475–91.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences, 6*(6), 237–242.

Cullen, K. E. (2004). Sensory signals during active versus passive movement. *Current Opinion in Neurobiology, 14*, 698–706.

Deecke, L., Becker, W., Grözinger, B., Scheid, P., & Kornhuber, H. H. (1973). Human brain potentials preceding voluntary limb movements. In W. C. McCallum & J. R. Knott (Eds.), *Electroencephalography and clinical neurophysiological supplement: Event-related slow potentials of the brain: Their relations to behavior* (Vol. 33, pp. 87–94). Elsevier: Amsterdam.

Desantis, A., Roussel, C., & Waszak, F. (2011). On the influence of causal beliefs on the feeling of agency. *Consciousness and Cognition, 20*(4), 1211–1220.

Desantis, A., Weiss, C., Schutz-Bosbach, S., & Waszak, F. (2012). Believing and perceiving: Authorship belief modulates sensory attenuation. *PLoS ONE, 7*(5), e37959.

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: A positron emission tomography study. *Neuroimage, 18*, 324–333.

Feltz, A., & Cokely, E. (2009). Do judgments about freedom and responsibility depend on who you are?: Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition, 18*, 342–50.

Fourneret, P., & Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia, 36*(11), 1133–1140.

Frith, C., Blakemore, S., & Wolpert, D. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society B, 355*(1404), 1771–1788.

Gallagher, S. (2007). The natural philosophy of agency. *Philosophy Compass, 2*(2), 347–357.

Gomes, G. (1999). Volition and the readiness potential. *Journal of Consciousness Studies, 6*(8–9), 59–76.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*, 128–133.

Haggard, P., & Tsakiris, M. (2009). The experience of agency: Feeling, judgment and responsibility. *Current Directions in Psychological Science, 18*(4), 242–246.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382–385.

Horgan, T. (2010). The phenomenology of agency and the Libet results. In W. Sinnott-Armstrong & L. Nadel (Eds.), *Conscious will and responsibility: A tribute to Benjamin Libet*. New York: Oxford University Press.

Horgan, T., Tienson, J., & Graham, G. (2003). The phenomenology of first-person agency. In S. Walter & H.-D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 323–40). Exeter: Imprint Academic.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*, 309–324.

Knoblich, G., & Kircher, T. T. J. (2004). Deceiving oneself about being in control: Conscious detection of changes in visuomotor coupling. *Journal of Experimental Psychology-Human Perception and Performance, 30*(4), 657–666.

Levy, N. (2005). Libet's impossible demand. *Journal of Consciousness Studies, 12*, 67–76.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*, 529–566.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain, 106*, 623–642.

Malle, B., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101–121.

McCallum, W. C. (1988). Potentials related to expectancy, preparation and motor activity. In T. W. Picton (Ed.), *Human event-related potentials* (EEG handbook, Vol. 3, pp. 427–534). Amsterdam: Elsevier.

Mele, A. (2009). *Effective intentions: The power of conscious will*. New York: Oxford University Press.

Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology: General, 136*(2), 184–199.

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition, 21*(1), 68–59.

Moore, J. W., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition, 17*(1), 136–144.

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition, 18*, 1056–64.

Nahmias, E. (2002). When consciousness matters: A critical review of Daniel Wegner's. *The Illusion of Conscious Will, Philosophical Psychology, 15*(4), 527–41.

Nahmias, E. (2013). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Freedom and responsibility, Vol. 4). Cambridge, MA: MIT Press.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies, 11*(7–8), 162–79.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561–584.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind and Language, 19*, 473–502.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science, 331*(6023), 1401–3.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk Intuitions. *Nous, 41*, 663–685.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*(1), 179–217.

Rigoni, D., Kühn, S., Sartori, G., & Brass, M. (2011). Inducing disbelief in free will alters brain correlates of preconscious motor preparation: The brain minds whether we believe in free will or not. *Psychological Science, 22*(5), 613–8.

Roskies, A. (2010a). How does neuroscience affect our conception of volition? *Annual Review of Neuroscience, 33*, 109–30.

Roskies, A. (2010b). Why Libet's studies don't pose a threat to free will. In W. Sinnott-Armstrong & L. Nadel (Eds.), *Conscious will and responsibility* (pp. 11–22). New York: Oxford University Press.

Sato, A. (2009). Both motor prediction and conceptual congruency between preview and action-effect contribute to explicit judgment of agency. *Cognition, 110*(1), 74–83.

Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: Discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition, 94*(3), 241–255.

Schlegel et al. (2013). Barking up the wrong free: readiness potentials reflect processes independent of conscious will. *Experimental Brain Research*. DOI:10/1007/s00221-013-3479-3.

Schurger, A., Sitt, J., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences, 109*, E2904–E2913.

Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Frontiers of science and philosophy* (pp. 35–78). Pittsburgh: University of Pittsburgh Press.

Sinnott-Armstrong, W., & Nadel, L. (2010) Conscious will and responsibility. Cambridge, MA: MIT Press.

Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass, 5*(2), 199–212.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multi-factorial two-step account of agency. *Consciousness and Cognition, 17*(1), 219–239.

Trevena, J. A., & Miller, J. (2002). Cortical movement preparation before and after a conscious decision to move. *Consciousness and Cognition, 11*, 162–90.

Tsakiris, E., & Haggard, P. (2003). Awareness of somatic events associated with a voluntary action. *Experimental Brain Research, 149*(4), 439–446.

Turner, J., & Nahmias, E. (2006). Are the folk agent-causationists? *Mind and Language, 21*(5), 597–609.

Wakefield, J., & Dreyfus, H. (1991). Intentionality and the phenomenology of action. In E. Lepore & R. van Gulick (Eds.), *John Searle and his critics* (pp. 259–70). Oxford: Blackwell.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist, 54*, 480–491.

Wegner, D. M., Fuller, V. A., & Sparrow, B. (2003). Clever hands: Uncontrolled intelligence in facilitated communication. *Journal of Personality and Social Psychology, 85*(1), 5–19.

Determinism and Its Relevance to the Free-Will Question

15

Mark Balaguer

Contents

Abstract

This paper begins with an argument for the claim that the compatibilism question (i.e., the question of whether free will is compatible with determinism) is less relevant than it might seem to questions about the metaphysical nature of human decision-making processes. Next, libertarianism (i.e., the view that human beings possess an indeterministic, libertarian sort of free will) is defended against a number of objections, and it is argued that there's a certain subset of our decisions (which can be called *torn* decisions) for which the following is true: If these decisions are appropriately undetermined at the moment of choice, then they are also free in a libertarian sort of way. This is an extremely important

M. Balaguer
Department of Philosophy, California State University, Los Angeles, CA, USA
e-mail: mbalagu@exchange.calstatela.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_121,

and surprising result; it entails that the question of whether libertarianism is true reduces to the straightforwardly empirical question of whether our torn decisions are in fact undetermined (in the appropriate way) at the moment of choice. Finally, the paper ends by arguing that as of right now, there is no compelling empirical evidence on either side of this question. In other words, the question of whether our torn decisions are appropriately undetermined is an open empirical question. And from this, it follows that the question of whether libertarianism is true is also an open empirical question.

Introduction

I will do two things in this paper. Ultimately, I will discuss how much evidence there is for various kinds of determinism that might be relevant to the question of whether we humans have free will. Before I do that, however, I will discuss the issue of whether the determinism question is even relevant to questions about the kinds of freedom we have.

Is the Determinism Question Relevant to the Free-Will Question?

Let *determinism* (or as I will sometimes call it, *universal determinism*) be the thesis that every event is causally necessitated by prior events together with causal laws. *Prima facie*, this thesis seems to be incompatible with the thesis that human beings have free will. For if it was already causally determined a billion years ago that my life would take the exact course that it's actually taken – that all of my actions and decisions would turn out exactly as they have – then it would seem to follow that I don't have free will. But some philosophers deny this; they endorse *compatibilism*, i.e., the thesis that free will is compatible with determinism. Compatibilists usually try to motivate their view by (a) providing an analysis of the notion of free will and (b) arguing that, given their analysis, it follows that free will is compatible with determinism. For instance, Hume famously argued that free will is compatible with determinism because free will is essentially just *the ability to do what you want*. Hume put his analysis like this (1748, p. 104):

> By liberty, then, we can only mean *a power of acting or not acting according to the determinations of the will*; that is, if we choose to remain at rest, we may; if we choose to move, we also may.

Putting this into contemporary lingo (and altering it somewhat), we arrive at the following definition:

> A person S is *Hume-free* iff S is capable of acting in accordance with his or her choices and of choosing in accordance with his or her desires; i.e., iff it is the case that if he or she chooses to do something, then he or she does it, and if (all things considered) he or she wants to make some choice, then he or she does make that choice.

Hume's argument for compatibilism is based on the following two claims:

(i) Hume-freedom captures the ordinary notion of free will (that is, Hume-freedom *is* free will).

(ii) Hume-freedom is compatible with determinism.

The only controversial claim here is (i). In other words, (ii) is entirely obvious. Hume-freedom requires only that our actions flow out of our decisions and our decisions flow out of our desires, but this could be the case even if all of our desires and decisions and actions are causally determined. To appreciate this, it's sufficient to notice that it could be that (a) it was causally determined by the big bang and the laws of nature that we would all have the desires that we actually do have and (b) our desires causally determine our decisions and our decisions causally determine our actions.

Given this, incompatibilists (i.e., those who think that free will is incompatible with determinism) are forced to reject premise (i) of the Humean argument. And that's exactly what they do; they reject the Humean analysis of free will, and they endorse instead a *libertarian* analysis. There are a few different ways to define libertarian-freedom (or as we can also call it, *L-freedom*), but one way to do this is as follows:

> A person is *libertarian-free* (or *L-free*) if and only if he or she makes at least some decisions that are such that (a) they are both undetermined and appropriately nonrandom and (b) the indeterminacy is relevant to the appropriate nonrandomness in the sense that it *generates* the nonrandomness, or *procures* it, or *enhances* it, or *increases* it, or something along these lines.

A lot needs to be said about what appropriate nonrandomness consists in, but the basic idea is that in order for a decision to count as appropriately nonrandom, the agent in question has to be centrally involved in the decision. Different philosophers might say slightly different things about what exactly this amounts to, but I think most libertarians would say that the most important requirement for appropriate nonrandomness is that the agent in question has to *author and control* the decision; i.e., it has to be *her* decision, and she has to control which option is chosen. (Other requirements for appropriate nonrandomness might include some kind of *rationality* and what Kane (1996) calls *plural* authorship, control, and rationality; but none of this will matter in what follows.)

In any event, the main point here is that incompatibilists disagree with Humean compatibilists about what free will *is*. Both parties can agree that Hume-freedom is compatible with determinism and L-freedom isn't. But they disagree about whether free will is Hume-freedom or L-freedom.

It's also important to note that the Humean analysis of free will isn't the only compatibilist analysis in the literature. Alternative compatibilist analyses (of not just free will, but moral responsibility) have been put forward by a number of different philosophers, e.g., P.F. Strawson (1962); Frankfurt (1971); Watson (1975); Wolf (1990); Fischer (1994); Wallace (1994); Mele (1995); and McKenna (2012), to name just a few. And, of course, each different analysis of free will gives us a different kind of freedom; thus, e.g., we can define *Frankfurt-freedom*, *Watson-freedom*, *Fischer-freedom*; and so on.

So the question of whether free will is compatible with determinism essentially boils down to the question of whether free will is L-freedom or one of these compatibilist kinds of freedom.[1] Now, this might seem pretty straightforward, but the question is notoriously difficult, and there's no consensus on what the right answer is. There are numerous arguments on both sides of the debate – e.g., there's the Frankfurt-case argument for compatibilism, first articulated in Frankfurt (1969); there's the consequence argument for incompatibilism (the locus classicus of this argument is van Inwagen (1975), but see also Ginet (1966) and Wiggins (1973)); there's the manipulation argument for incompatibilism (see, e.g., Pereboom (2001)) – but none of these arguments have proved really compelling, and at the present time, it seems fair to say that, as a community, we just don't know whether free will is compatible with determinism.

Given this, you might think we should conclude that we don't know the answer to the main question of the present section – i.e., that we don't know whether

The determinism question: Is determinism true?

is relevant to

The do-we-have-free-will question: Do humans have free will?

More specifically, you might think we should say that in order to figure out whether the determinism question is relevant to the do-we-have-free-will question, we first need to figure out what free will is. If free will is L-freedom, then the determinism question *is* relevant to the do-we-have-free-will question, and if free will is some compatibilist kind of freedom, then the determinism question *isn't* relevant to the do-we-have-free-will question.

There's a sense in which this is right, but it seems to me that it's also misleading. For there's another way to think about the issues here, and on this other way of conceptualizing things, less turns on the compatibilism debate than the above remarks suggest. In particular, it seems to me that regardless of whether compatibilism is true, we already know right now that the determinism question is highly relevant to an important question about the nature of human freedom. To appreciate this, notice first that the do-we-have-free-will reduces to (or is subsumed by, or some such thing) the following two more fundamental questions:

The what-is-free-will question: What is free will? That is, is it Hume-freedom, or L-freedom, or what?

The which-kinds-of-freedom-do-we-have question: Which kinds of freedom do humans have? That is, do they have L-freedom?; and do they have Hume-freedom?; and do they have Frankfurt-freedom?; and so on. (Actually, to be more precise, we can formulate this question as asking which kinds of "freedom" humans have, since some or all of the kinds of "freedom" we're asking about

[1]You might think that for some of the so-called compatibilist kinds of freedom in the literature, it's actually not obvious that they really are compatible with determinism. If this is right, then if it also turned out that one of these kinds of freedom was free will, then we couldn't settle the compatibilism question by merely determining that free will was the given kind of freedom; we would also need to determine whether the given kind of freedom was compatible with determinism.

here might fail to *be* free will, according to the correct answer to the what-is-free-will question.)

We can think of the latter question here (i.e., the which-kinds-of-freedom-do-we-have question) as the *metaphysical* part of the do-we-have-free-will question – i.e., the part that's actually about the nature of human beings and human decision-making processes. The former question isn't really about *us* at all; it's a *semantic* question.[2] But notice now that the determinism question is obviously relevant to the which-kinds-of-freedom-do-we-have question because it's relevant to the *libertarian question*, i.e., the question of whether humans are L-free. (Actually, one might deny that the determinism question is relevant to the libertarian question; I'll discuss this in the next section, but for now, I will ignore this.) In any event, this is what I had in mind when I said that there's another way to think about the issues here. For if we assume that the determinism question is indeed relevant to the libertarian question, then without even answering the compatibilism question (or the what-is-free-will question), we arrive at the result that the determinism question is relevant to an interesting and important question about the nature of human freedom, namely, the libertarian question, i.e., the question of whether we're L-free.

Now, I suppose you might claim that if it turns out that L-freedom isn't free will (i.e., that L-freedom isn't the referent of the ordinary term "free will"), or if it turns out that L-freedom isn't required for moral responsibility, then the libertarian question is, in fact, *not* interesting or important. But this is just false. The question of whether we possess an indeterministic, libertarian sort of freedom is *intrinsically* interesting and important. In other words, even if L-freedom isn't required for moral responsibility, and even if it isn't the referent of the ordinary term "free will," the question of whether we actually *possess* L-freedom is itself an interesting and important question about the nature of human beings and human decision-making processes. (Of course, there are *some* people who *aren't* interested in the question of whether we're L-free, but so what? – that's true of every question. There are lots of people who aren't interested in whether Alpha Centauri has planets, or what the ultimate laws of physics are, or whether humans are morally responsible for their actions. That doesn't make these questions uninteresting or unimportant.) In any event, since the libertarian question is itself an interesting and important question about the nature of human freedom, it follows that, regardless of whether compatibilism is true, the determinism question is relevant to an interesting and important question about human freedom, namely, the libertarian question, i.e., the question of whether we're L-free.

---

[2]Some people would say that the what-is-free-will question is essentially equivalent to the question, "Which kind(s) of freedom are required for moral responsibility?" But (a) I think it can be argued that the which-kinds-of-freedom-are-required-for-moral-responsibility question is *also* a semantic question (because it's just a subquestion of the what-is-moral-responsibility question), and (b) even if it's not a semantic question, it's pretty clearly not about the metaphysical nature of human decision-making processes.

Is the Determinism Question Relevant to the Libertarian Question?

Libertarianism is often defined as the view that (a) human beings possess L-freedom and (b) L-freedom is free will. But in what follows, I will be concerned with thesis (a) only, and so I will take libertarianism to be the view that humans are L-free, and when I talk about "the libertarian question," I will have in mind the question of whether we are L-free.

Now, *prima facie*, it seems obvious that the determinism question is relevant to the libertarian question. After all, the libertarian question is just the question of whether we're L-free, and L-freedom requires indeterminism, and so it seems obvious that the determinism question is relevant to the libertarian question. But you might question this. You might think we can know by logic alone that we're *not* L-free because the notion of L-freedom is *incoherent*. And if this is right, then the question of whether determinism is true is actually not relevant to the question of whether we're L-free.

The key claim here is obviously that libertarianism is incoherent. This point has been argued by a number of different philosophers – see, e.g., Hobbes (1651), Hume (1748), and Hobart (1934) – but the reasoning is always very similar. One way to formulate the argument here is as follows:

> *Anti-libertarian argument*: Any event that's undetermined is uncaused, and so it *just happens* – i.e., it happens randomly. Thus, if we insert an undetermined event into a decision-making process, we're inserting a *random* event into that process. But given this, it's hard to see how there could be any undetermined events in our decision-making processes that increase (or generate, or enhance, or whatever) appropriate nonrandomness. Appropriate nonrandomness has to do with the agent being in control. How could this be increased by inserting a random event into the process? It seems that it couldn't, and so it seems that libertarianism couldn't be true.

How libertarians respond to this argument is largely determined by the kind of libertarianism they endorse. Broadly speaking, there are three different kinds of libertarian views. First, there are *event-causal* views, which hold that undetermined L-free decisions are *nondeterministically caused* by prior events (I think the best way to fill this view in is to say that these decisions are *probabilistically caused* by prior events, most notably events having to do with the agent's reasons); event-causal views have been developed by, e.g., van Inwagen (1983), Kane (1996), Ekstrom (2000), and Balaguer (2010). Second, there are *noncausal* libertarian views, which hold that L-free choices are completely uncaused (see, e.g., Ginet (1990)). And third, there are *agent-causal* views, which hold that L-free decisions are caused but not by prior events; rather, they're directly caused by *persons* via a special causal relation known as *agent causation*; views of this kind have been endorsed by, e.g., Reid (1788), Chisholm (1964), R. Taylor (1966), C.A. Campbell (1967), Thorp (1980), Rowe (1987), O'Connor (2000), and Clarke (1993).

In this section, I will briefly sketch an event-causal response to the above anti-libertarian argument. Most libertarians who have tried to respond to the anti-libertarian argument have done so by trying to explain how our decisions could be simultaneously undetermined and appropriately nonrandom. But I think

libertarians can motivate a much stronger claim than this. I think they can motivate the following thesis:

> (L) There's an important subset of our decisions (I'll call them *torn decisions*, and I'll characterize them shortly) for which the following is true: If they're undetermined in the right way, then they're also appropriately nonrandom (i.e., we author and control them), and the indeterminacy in question procures the nonrandomness, and so they're L-free.

This is a really strong result; if it's right, what it shows is that the anti-libertarian argument gets things exactly backwards; more precisely, it shows that a certain kind of indeterminism is sufficient for the truth of libertarianism. And, of course, it also shows that, contrary to what the anti-libertarian argument suggests, the determinism question is definitely relevant to the libertarian question.

Before I argue for (L), I first need to say what a torn decision is, and I need to say what the relevant sort of indeterminacy is, i.e., I need to say exactly how a torn decision needs to be undetermined in order to be L-free. Thus, let me start by saying this:

> A *torn decision* is a decision in which the person in question has reasons for multiple options, feels torn as to which option is best, and decides without resolving the conflict, i.e., decides while feeling torn.

I think we make decisions like this several times a day about things like whether to have eggs or cereal for breakfast, and whether to drive or bike to the office, and so on. But we can also make torn decisions in connection with big life-changing decisions; e.g., you might have a good job offer in a bad city, and you might have a deadline that forces you to decide while feeling utterly torn. (Torn decisions are obviously a lot like Kane's self-forming actions, or SFAs. But there are a few differences. Note, in particular, that unlike SFAs, torn decisions are not defined as being undetermined. They're defined in terms of their phenomenology. Thus, we know from experience that we do make torn decisions, and it's an empirical question whether any of these decisions are undetermined in the right way.)

Next, let me define the relevant sort of indeterminacy, i.e., the sort that's needed for an ordinary torn decision to be fully L-free. We can do this as follows:

> A torn decision is *wholly undetermined* at the moment of choice – or, as I'll also say, *TDW-undetermined* – iff the actual, objective moment-of-choice probabilities of the various reasons-based tied-for-best options being chosen match the reasons-based probabilities (or the phenomenological probabilities), so that these moment-of-choice probabilities are all roughly even, given the complete state of the world and all the laws of nature, and the choice occurs without any further causal input, i.e., without anything else being significantly causally relevant to which option is chosen.

Given this, we can say that *TDW-indeterminism* is the view that some of our torn decisions are TDW-undetermined. And now, given all of these definitions, I can reformulate thesis (L) as follows:

> *Central-Libertarian-Thesis*: If our torn decisions are undetermined in the right way – i.e., if they're wholly undetermined, or TDW-undetermined – then we author and control them, and they're appropriately nonrandom and L-free. Or more succinctly: *If TDW-indeterminism is true, then libertarianism is true.*

I argued for this thesis at length in a recent book (2010). I can't rehearse the whole argument here, but I'd like to provide brief formulations of two (related) arguments for Central-Libertarian-Thesis.

The first argument is easier to articulate if we assume a weak, token-token mind-brain identity theory – or more precisely, if we assume that ordinary human decisions are neural events. I don't actually need this assumption, but it makes the argument run more smoothly. In any event, given this background assumption, let's look at an example of a torn decision. Suppose that Ralph has to choose between two options, O and P, and suppose that he makes a torn decision to go with O rather than P. The details don't matter; option O could be something important like a new job, or it could be something trivial like a chocolate ice cream cone. All that matters is that Ralph makes a conscious torn decision to go with option O. Given this, if we assume that Ralph's decision was TDW-undetermined, then we get the following results:

A. Ralph's choice was conscious, intentional, and purposeful, with an actish phenomenology – in short, it *was* a Ralph-consciously-choosing event, or a Ralph-consciously-doing event (we actually know all of this independently of whether the choice was TDW-undetermined).
B. The choice flowed out of Ralph's conscious reasons and thought in a nondeterministically event-causal way.
C. Nothing external to Ralph's conscious reasons and thought had any significant causal influence (at the moment of choice – i.e., after he moved into a torn state and was about to make his decision) over how he chose, so that the conscious choice itself *was* the event that settled which option was chosen. (If you like, we can put it this way: The conscious choice itself *was* the undetermined physical event that settled which option was chosen.)

My first argument for Central-Libertarian-Thesis is based on the observation that, given (A)–(C), it seems to make sense to say that Ralph authored and controlled his decision. For (A)–(C) seem to give us the twofold result that (i) *Ralph did it and* (ii) *nothing made him do it*; and, intuitively, it seems that (i) and (ii) are sufficient for authorship and control.

Now, to get the result that Ralph's decision is appropriately nonrandom and L-free, we also need to argue that (a) his decision satisfies the other conditions for appropriate nonrandomness, aside from authorship and control (i.e., rationality, the plurality conditions, and so on), and (b) the fact that Ralph's decision is TDW-undetermined *procures* the result that it's appropriately nonrandom and L-free. Point (a) is actually very easy to argue for; I don't have the space to get into this here, but see Balaguer (2010). Point (b), on the other hand, should already be clear from the above argument; for the fact that Ralph's decision was TDW-undetermined played a crucial role in the argument for the claim that he authored and controlled the decision. It's *because* the decision was TDW-undetermined that we get the result that nothing made Ralph choose O over P. Now, it's important to note that the idea here isn't that TDW-indeterminacy actively *generates* authorship and control; the idea is rather that it *blocks a destroyer* of authorship and control. The destroyer of authorship and control would be a moment-of-choice causal influence from something external to the

agent's conscious reasons and thought. But TDW-indeterminacy rules out the possibility of such a destroyer – if a torn decision is TDW-undetermined, then at the moment of choice, nothing external to the agent's conscious reasons and thought comes in and causally influences which option is chosen – and this is why TDW-indeterminacy can be seen as *procuring* authorship and control.

My second argument for Central-Libertarian-Thesis is based on the fact that when we make torn decisions, it *feels* as if we author and control them. The argument can be put like this:

1. The only initially plausible reason to doubt the phenomenology of our torn decisions – i.e., the only reason to doubt our feeling that we author and control these decisions – is that it might be that, unbeknownst to us, our torn decisions are causally influenced (at the moment of choice) by events that are external to our conscious reasons and thought in a way that's inconsistent with the idea that we author and control these decisions. (For example, it could be that our torn decisions are deterministically caused by wholly non-mental brain events that precede our torn decisions in our heads.) But
2. If our torn decisions are TDW-undetermined, then they're *not* causally influenced (at the moment of choice) by anything external to our conscious reasons and thought. Thus,
3. The assumption that our torn decisions are TDW-undetermined seems to eliminate the only initially plausible worry we might have about the accuracy of the phenomenology of our torn decisions. Therefore, it seems plausible to conclude that
4. *If* our torn decisions are TDW-undetermined, then the phenomenology of our torn decisions is accurate and, hence, we author and control these decisions; moreover, it should be clear that the TDW-indeterminacy is *procuring* the authorship and control here, and so we get the result that if our torn decisions are TDW-undetermined, then they're also appropriately nonrandom and L-free.[3] In other words, we get the result that
5. Central-Libertarian-Thesis is true.

The two arguments for Central-Libertarian-Thesis that I just articulated are obviously very quick, and there are a number of different worries that one might have about them. I won't be able to respond to all of these worries here, but I'd like to say a few words about two of them. I'll start with this one:

> *The Rollback Objection*: Suppose that Ralph is torn between two options, O and P, and eventually chooses O in a torn decision sort of way. And now suppose that God "rolls back" the universe and "replays" the decision. If the decision is undetermined at the moment of choice, then it seems that the decision might very well go differently the second time around, even if everything about the past – in particular, everything about Ralph and his reasons – remained the same. Indeed, if the decision is TDW-undetermined, then it seems that if God "played" the decision 100 times, we should expect that Ralph would choose

[3]Actually, to fully motivate (4), we would also need to argue that if our torn decisions are TDW-undetermined, then they satisfy the other requirements for appropriate nonrandomness, i.e., rationality and the plurality conditions. But, again, this point is easy to argue; see Balaguer (2010), Sects. 3.3.4–3.3.5.

O and P about 50 times each. But given this – given that Ralph would choose differently in different "plays" of the decision, without *anything* about his psychology changing – it's hard to see how we can maintain that Ralph authored and controlled the decision. It seems to be a matter of *chance* or *luck* what he chose, and to the extent that this is right, it seems that Ralph didn't author or control the choice.

The first point I want to make in response to this objection is that it simply doesn't follow from the fact that Ralph would choose differently in different "plays" of the decision that he didn't author or control the decision. There is no inconsistency in claiming that (a) Ralph chooses differently in different plays of the decision and (b) in each of the different plays of the decision, it is *Ralph* who does the choosing and who authors and controls the choice. Indeed, given that Ralph is making a *torn* decision, the hypothesis that it's *him* who's making the decision (or who's authoring and controlling the decision) seems to *predict* that he would choose differently in different plays of the decision. It would seem very suspicious if he always chose the same option in the various plays of the decision; in that scenario, it would be plausible to think: "That can't be a coincidence; something must be *causing* him to choose that way; and since (by assumption) his conscious reasons and thought aren't causing this, it must be something else, e.g., a random, non-mental event in his nervous system, or a subconscious mental state." But if Ralph chose *differently* in different plays of the decision, that would fit perfectly with the hypothesis that the choice is flowing from *him*, or from his conscious reasons and thought; for since Ralph is making a torn decision, we know by assumption that he is neutral between his two live options, at least in his conscious thought. Thus, it seems to me that since Ralph is torn, if he chose differently in different plays of the universe, that wouldn't undermine the hypothesis that he authors and controls the decision; on the contrary, it would *confirm* that hypothesis.

(You might think that if there was a fixed probability that Ralph was going to choose option O – or, more specifically, if there was a 0.5 probability that he would choose O and a 0.5 probability that he would choose P – then it was just a matter of luck that he in fact *did* choose O, and so it couldn't be that he authored and controlled the decision. My response to this is that if Ralph's decision was TDW-undetermined, then (i) it was *his reasons* that caused it to be the case that the probabilities in question were 0.5, and (ii) despite the fact that there were fixed probabilities here, it is still true that the choosing of O over P was done *by Ralph*, because the event in question *was* a Ralph-consciously-choosing event, and this event wasn't causally influenced by anything external to Ralph's conscious reasons and thought.)

The second objection I want to consider here is the following:

The Agent-Causal Objection: The notion of control that you're working with (and the notion of authorship too, but let's just focus on the case of control) is too weak. Something more substantive is needed for control. In particular, it seems that something like agent causation is needed. In other words, when someone makes a torn decision, in order for it to be the case that the agent in question controls which option is chosen, it needs to be the case that he or she *causes* the given option to be chosen.[4]

[4]Pereboom raises a worry like this about my view in his (forthcoming).

I think it can be argued that agent causation just *isn't* needed for authorship and control, but I won't try to argue this point here. Rather, I want to argue that in the present context, the question of whether agent causation is required for authorship and control doesn't really matter. To bring this out, let me start by distinguishing two different kinds of control – *causal-control* and *noncausal-control* – where the former requires agent causation (or something like it) and the latter doesn't. I won't try to give precise definitions of these two notions; all I'll say (and all that will matter here) is that when I use the term "noncausal-control," I'm talking about a kind of control that applies to ordinary torn decisions if they're TDW-undetermined; i.e., it applies to torn decisions like Ralph's, where the agent makes a conscious decision with an actish phenomenology and which option is chosen isn't significantly causally influenced (at the moment of choice) by anything external to the agent's conscious reasons and thought, so that the conscious choice itself *is* the event that settles which option is chosen. Beyond this (and beyond the fact that causal-control requires agent causation and noncausal-control doesn't), it won't matter how exactly these two kinds of control are defined. But for the sake of argument, let's pretend that we've got two precisely defined kinds of control here. Given this, one question we might ask is the following:

> *The what-is-control question*: What is control? That is, which of the various kinds of control that we find in the literature is *real* control? Is causal-control real control? Is noncausal-control? Are both? Is neither?

But why should libertarians care about this question? They don't need to claim that if our torn decisions are TDW-undetermined, then they're authored and controlled by us and L-free in the only senses of these terms that anyone might care about, or in the senses of these terms that philosophers have traditionally cared about. All they need is this:

> (*) If our torn decisions are TDW-undetermined, then they're authored and controlled by us and appropriately nonrandom and L-free in interesting and important ways that are worth wanting and worth arguing for and that libertarians can hold up and say, "This gives us a noteworthy kind of libertarian free will."

Now, don't take me to be saying more than I am here. I'm not saying that libertarians can define authorship and control and L-freedom *however they want to*; they can't just define these terms in ridiculously weak ways and then claim victory. I don't need to argue that the kind of L-freedom I've articulated – the kind that we get if our torn decisions are TDW-undetermined (i.e., the kind that involves noncausal-control) – is the one and only kind of L-freedom that anyone might care about. But I do need it to be the case that this kind of L-freedom is interesting, worth wanting, worth arguing for, and so on. In other words, I need (*).

But I think the above arguments for Central-Libertarian-Thesis do motivate (*). Let's return to Ralph's decision. If it's TDW-undetermined, then (a) the choice was conscious, intentional, and purposeful, with an actish phenomenology – in short,

it *was* a Ralph-consciously-choosing event, or a Ralph-consciously-doing event; and (b) the choice flowed out of Ralph's conscious reasons and thought in a nondeterministically event-causal way; and (c) nothing external to Ralph's conscious reasons and thought had any significant causal influence (after he moved into a torn state and was about to make his decision) over how he chose, so that the conscious choice itself *was* the event that settled which option was chosen. This might not give us every kind of L-freedom you might have wanted, but it clearly gives us *one important kind* of L-freedom – a kind that libertarians can hang their hats on and that's worth wanting and arguing for and so on. After all, in this scenario, the event that settles which option is chosen *is* the conscious decision – i.e., it's the event with a me-consciously-choosing-now phenomenology.

There is obviously a lot more to say about all of this. In Balaguer (2010), I develop the arguments for Central-Libertarian-Thesis a lot more thoroughly, and I respond to a number of different objections to these arguments. For instance, I say more by way of response to the luck objection; and I respond to the worry that the kind of L-freedom I've been describing here isn't robust enough to give us moral responsibility; and I also respond to the worry that this kind of L-freedom isn't worth wanting because torn decisions are trivial. Unfortunately, though, I don't have the space to discuss these issues here.

In any event, if what I've argued in this section is correct, then we have the result that if TDW-indeterminism is true (i.e., if some of our torn decisions are TDW-undetermined), then libertarianism is also true (i.e., humans are L-free). But given this, it follows pretty quickly that the question of whether libertarianism is true just reduces to the question of whether TDW-indeterminism is true.[5] And if this is right, then we have an answer to the main question of this section: The determinism question is definitely relevant to the libertarian question.

[5]To argue for this, libertarians need to argue that if TDW-indeterminism *isn't* true, then libertarianism isn't true either – i.e., that if our torn decisions aren't TDW-undetermined, then we aren't L-free. Now, you might doubt this, because you might think that even if our torn decisions aren't L-free, some of our non-torn decisions could be L-free. But it's pretty easy to argue – and I do argue for this point in Balaguer (2010) – that if none of our torn decisions is L-free, then it's very likely that we don't make any L-free choices at all. The more important worry about the above thesis (i.e., the thesis that if TDW-indeterminism isn't true, then libertarianism isn't true either) is that even if our torn decisions aren't TDW-undetermined (i.e., even if they aren't *wholly* undetermined), they could still be *partially* undetermined in a certain way. To say that a torn decision is partially undetermined in the sense I have in mind here – or what comes to the same thing, partially determined – is to say that, at the moment of choice, factors external to the agent's conscious reasons and thought causally influence (but don't causally determine) which tied-for-best option is chosen. I think it can be argued that if our torn decisions are partially undetermined in this way, then they're also *partially L-free*. Thus, to be precise, what we need to say here is not that if TDW-indeterminism isn't true, then we aren't L-free, but that if TDW-indeterminism isn't true, then we aren't *fully* L-free. And so to get the result that if TDW-indeterminism isn't true, then libertarianism isn't true, we need to be clear that we're defining libertarianism as the view that humans are *fully* L-free.

Is There Any Good Reason to Doubt that Our Torn Decisions Are TDW-Undetermined?

We found in the last section that the question of whether libertarianism is true boils down to the question of whether TDW-indeterminism is true (i.e., the question of whether some of our torn decisions are TDW-undetermined). In this section, I want to discuss the question of whether we have any good reason to reject TDW-indeterminism. I think it's pretty obvious that, at present, we don't have any good reason to *endorse* TDW-indeterminism, but you might think we have good reason to *reject* it, because you might think we have good reason to endorse some deterministic thesis that's incompatible with TDW-indeterminism. I will argue in this section, however, that as of right now, we don't have any good reason to believe any such deterministic thesis.

Is There Any Good Reason to Believe Universal Determinism?

Let *universal determinism* (or UD) be the thesis that I've been calling "determinism" – i.e., the thesis that every event is causally necessitated by prior events together with causal laws. It's pretty easy to see that as of right now, we don't have any good reason to believe UD. For UD is true only if all quantum events are determined, and as of right now, we don't have any good reason to believe that all quantum events are determined. The first point to be made here is that quantum mechanics (or QM) contains probabilistic laws; it tells us, for instance, that if an electron is spin-up in a particular direction x, then it's in a superposition state with respect to its spin in the orthogonal direction y, and if we measure it for spin in the y-direction, then it will collapse into either a spin-up state or a spin-down state, and there's a 0.5 probability that it will collapse into a spin-up state and a 0.5 probability that it will collapse into a spin-down state.

Now, of course, the fact that QM contains probabilistic laws of this kind does not by itself give us reason to doubt UD; for it could be that there are hidden facts (or as physicists say, *hidden variables*) about electrons that are spin-up in the x-direction that determine whether they will collapse into a spin-up state or a spin-down state when measured for spin in the y-direction. But the problem is that there is no good evidence for the existence of hidden variables of this kind, and so, for all we know, it could just as easily be that when electrons that are spin-up in the x-direction are measured for spin in the y-direction, *nothing* determines whether they collapse into a spin-up state or a spin-down state; in other words, for all we know, it could be that events of this kind – i.e., events involving quantum wave-function collapses – are genuinely undetermined.

This is not to say that we have good reason to endorse an indeterministic view of these events. Rather, it's to say that as of right now, we have no good reason to reject an indeterministic view. In other words, the question of whether these quantum collapse events are genuinely undetermined or just apparently undetermined (i.e., really determined) is an open question. There is simply no good evidence on either side of the debate. Or to put the point differently, there's

no good evidence for any (deterministic or indeterministic) interpretation of QM. An interpretation of QM is essentially a theory of what's going on in quantum collapse events of the above kind; there are numerous interpretations in the literature, some deterministic and some indeterministic, but at present, there isn't any evidence for any of them, and more generally, there isn't any compelling reason to endorse a deterministic or an indeterministic view of quantum collapse events.[6]

Is There Any Good Reason to Believe Macro-Level Determinism?

If the arguments of the previous section are cogent, then there's no good reason to believe universal determinism. But in order to undermine TDW-indeterminism and libertarianism, you don't need to motivate universal determinism. Since TDW-indeterminism is about torn decisions only, you could undermine this thesis by arguing for the much weaker claim that all torn decisions are determined. One way to do this would be to point out that torn decisions are macro-level events and then argue for *macro-level determinism*, i.e., the view that all macro-level events are determined. Or, alternatively, you could undermine TDW-indeterminism by arguing for what might be called *virtual macro-level determinism*, i.e., the view that while it may be that some macro-level events are strictly undetermined (because they're composed of micro-level events, some of which are undetermined), it's also true that all macro-level events are, if not determined, at least virtually determined (where an event is *virtually determined* iff prior circumstances together with causal laws made it overwhelmingly likely that the given event would occur). In other words, the idea here is that while there may be some micro-level indeterminacies, these all "cancel out" before we get to the macro level, presumably because macro-level phenomena are composed of such large numbers of micro-level phenomena. (It should be clear that virtual macro-level determinism would undermine TDW-indeterminism; after all, it entails that all torn decisions are virtually determined, i.e., that for any torn decision, there's a unique option X such that prior events made it overwhelmingly likely that X would be chosen.)

The question I want to ask now is whether we have any good reason to believe macro-level determinism or virtual macro-level determinism. People sometimes claim that there's a good inductive argument for macro-level determinism (see, e.g., Honderich (2002)[7]). We might put the argument here as follows:

1. All the macro-level events that we have encountered have been causally determined by prior events together with causal laws. Therefore,
2. Macro-level determinism is true – i.e., all macro-level events are determined.

[6]Of course, there are people who favor certain interpretations over others, but there is pretty widespread agreement among those who work on the foundations of quantum mechanics that we do not have any solid evidence for any of the various interpretations and that when people embrace these interpretations, they are engaged in speculation.

[7]Actually, Honderich thinks we can use arguments like the one in the text to motivate not just macro-level determinism but universal determinism as well.

But this argument is misguided. In particular, premise (1) is unmotivated, controversial, and question begging. We encounter all sorts of macro-level events that, for all we know, could be undetermined – coin tosses, events in which a person contracts chicken pox from someone else, events in which macro-level measuring devices reveal quantum wave-function collapses, human decisions, chimp decisions, parakeet decisions, temper tantrums, etc. Now, of course, determinists have a story to tell about how it *could be* that events like these are deterministic; e.g., they can claim that if, say, Jack and Jill were both exposed to chicken pox and only Jack fell ill, this would not undermine determinism because it could be that there were hidden physical variables at work in the situation (e.g., factors having to do with the physical well-being of Jack and Jill, or the duration of their exposures, or whatever) that determined that Jack would contract the disease and Jill would not. And likewise for events of the other kinds listed above; determinists can say that events like coin tosses and decisions could be determined even if they don't seem determined to us, because it could be that there are hidden determining factors at work in such cases. I agree; for all we know, it *could be* that events of the above kinds are determined. But in the present context, this is entirely irrelevant. What advocates of the argument in (1)–(2) need isn't a story about how it *could be* that events of the above kinds are determined; what they need is a positive argument for the claim that, in fact, such events *are* determined.

But I take it that determinists don't have an argument of this kind. The argument they used to give here is that any apparently indeterministic behavior in macro-level systems must really be deterministic, because such systems are made up of micro-level systems whose behavior is deterministic. But this argument is no good, because (as we've seen) we currently have no more reason to believe micro-level determinism than macro-level determinism.

Now, I suppose one might respond here by claiming that every time we go looking for deterministic explanations, we find them. But this is just false. It's not just that we don't currently have deterministic explanations of events of the above kinds; it's that we haven't the foggiest idea how to proceed in trying to construct and justify such explanations.

The situation with virtual macro-level determinism is similar. One might try to argue for this view by saying something like the following:

1′. All the macro-level events that we've encountered have been either determined or virtually determined. Therefore,

2′. Virtual macro-level determinism is true – i.e., all macro-level events are either determined or virtually determined.

But this argument is flawed in the same way the (1)–(2) argument is flawed. In short, the problem is that (1′) is unmotivated, controversial, and question begging. There are lots of macro-level events – coin tosses, quantum-measurement events, decisions, and so on – that, for all we know, are neither determined nor virtually determined. In order for virtual macro-level determinists to motivate an inductive argument of the above kind, they would need to provide positive reasons for thinking that events like coin tosses and decisions and quantum measurements

are, in fact, either determined or virtually determined. But at present, there is simply no good reason to believe this.

Finally, it's worth noting here that if the remarks in this section are correct, they suggest not just that the above inductive arguments are noncogent, but that, right now, we don't have any good reason to believe macro-level determinism or virtual macro-level determinism.

Is There Any Good Reason to Believe Neural Determinism?

Since torn decisions are presumably neural events, you might think that we could undermine TDW-indeterminism (and hence libertarianism) by uncovering reasons to believe *neural determinism* (i.e., the view that all neural events are determined) or *virtual neural determinism* (i.e., the view that all neural events are either determined or virtually determined in the sense defined above). But, in fact, we don't have any good reason to believe either of these theses. If current neuroscientific theory were deterministic (or virtually deterministic), then we might be able to motivate neural determinism (or virtual neural determinism). But current neuroscientific theory is *not* deterministic (or virtually deterministic). Indeed, it treats a number of different neural processes probabilistically – e.g., synaptic transmission and spike firing. Consider, e.g., the following passages from a recent textbook on neuroscience (Dayan and Abbott 2001):

> I. . . . [synaptic] transmitter release is a stochastic process. Release of transmitter at a presynaptic terminal does not necessarily occur every time an action potential arrives and, conversely, spontaneous release can occur even in the absence of the depolarization due to an action potential. (p. 179)
>
> II. Because the sequence of action potentials generated by a given stimulus varies from trial to trial, neuronal responses are typically treated statistically or probabilistically. For example, they may be characterized by firing rates, rather than as specific spike sequences. (p. 9)

It is worth noting that some aspects of the indeterminacies in these processes (or the apparent indeterminacies, as the case may be) are caused by the indeterminacy (or apparent indeterminacy) in another process, namely, the opening and closing of ion channels. Now, to be sure, by treating these processes probabilistically, neuroscientists don't commit themselves to the thesis that, in the end, they are genuinely indeterministic. But the important point here is that they aren't committed to determinism either. The question of whether these processes are genuinely indeterministic simply isn't answered by neuroscientific theory. Indeed, it is a standard view among those who work in this area that for at least some of these processes (e.g., the opening and closing of ion channels), this isn't even a neuroscientific question, because it is already clear right now that there could not be deterministic neuroscientific explanations of the phenomena. In other words, the idea is that (a) from the point of view of neuroscience, these processes might as well be undetermined but (b) it *could* be that there are underlying

deterministic *physical* explanations of the phenomena. Thus, the question of whether there actually are such explanations is not a neuroscientific question at all; it is rather a question of physics, because the issue comes down to questions about the behavior of the elementary physical particles involved in the neural processes.

It sum, then, it seems to me that neuroscientific theory is neither deterministic nor virtually deterministic, and so it doesn't give us any reason to believe neural determinism or virtual neural determinism. And given this, it seems safe to conclude that as of right now, we don't have any good reason to believe neural determinism or virtual neural determinism.

Is There Any Good Reason to Believe Torn-Decision Determinism?

Finally, you might try to undermine TDW-indeterminism by arguing for *torn-decision determinism* (i.e., the view that all torn decisions are determined) or *virtual torn-decision determinism* (i.e., the view that all torn decisions are either determined or virtually determined). Or, of course, you could try to argue directly against TDW-indeterminism; i.e., you could try to give a direct argument for the claim that none of our torn decisions is TDW-undetermined. In this section, I will respond to one such argument, an argument based on the work of Benjamin Libet.

(It's worth noting that the argument based on Libet's work isn't the only argument against TDW-indeterminism that one might consider here. Another important argument – we might take it to be an argument for something like virtual torn-decision determinism – is based on Max Tegmark's (2000) argument for the claim that if there are any neural superposition states, they couldn't survive long enough to be affected by neural processes. One might also construct arguments against TDW-indeterminism by using considerations having to do with situationism (see, e.g., Isen and Levin 1972, Milgram 1969, and Nelkin 2005), the sluggishness of consciousness (see, e.g., Velmans 1991 and Wegner 2002), or the way in which humans are often out of touch with the real underlying reasons for their actions (see, e.g., Festinger 1957). In Balaguer (2010), I argue that none of these considerations provides us with any good reason to reject TDW-indeterminism; but unfortunately, I don't have the space to pursue any of this here.)

In any event, let's consider the argument against TDW-indeterminism that's based on Libet's work. Libet's studies were a follow-up to a neuroscientific discovery from the 1960s, in particular, the discovery that voluntary decisions are preceded in the brain by a slow change in electrical potential known as the *readiness potential* (see, e.g., Kornhuber and Deecke 1965). Libet's studies were an attempt to establish a timeline for the readiness potential, the conscious intention to act, and the act itself (see, e.g., Libet et al. 1983, and Libet 2002). His results suggest that the readiness potential appears about 350–400 milliseconds before the conscious intention to act and about 550 milliseconds before the act itself.

And given this, one might argue against TDW-indeterminism in something like the following way:

1. Conscious decisions are preceded by nonconscious brain processes (namely, the readiness potential) and are, in fact, nonconsciously initiated. Therefore, it seems likely that
2. Torn decisions are at least causally influenced by prior-to-choice nonconscious brain processes, and so they are not TDW-undetermined; indeed, they might be determined, or virtually determined, by prior-to-conscious-choice brain processes.

In other words, the idea here is that our torn decisions couldn't be TDW-undetermined because (to borrow Henrik Walter's (2001) phrasing) the "neural machinery" for starting our decisions is already up and running before our conscious thinking enters the picture.

One might try to attack the argument in (1)–(2) by questioning (1), but I won't pursue this strategy here. What I want to argue instead is that even if (1) is true, it does not give us any good reason to accept (2). The first point to note here is that we don't know what the *function* of the readiness potential is. In particular, it would be an unmotivated assumption to suppose that, in torn decisions, the readiness potential is part of a causal process that's relevant to which option is chosen. There are plenty of other things the readiness potential could be doing, aside from this. One way to appreciate this is to notice that libertarianism is perfectly consistent with the idea that various things involved with our torn decisions might be causally determined. In particular, a torn decision could be L-free even if it was determined in advance that (i) a torn decision would occur, and (ii) the choice would come from among the agent's reasons-based tied-for-best options, and (iii) the moment-of-choice probabilities of these options being chosen were all roughly even. The only thing that needs to be undetermined, in order for a torn decision to be L-free, is *which tied-for-best option is chosen*. Given this, here are two stories libertarians could tell about what the readiness potential could be doing (there are other stories as well – see, e.g., Mele (2009) – but these two will do):

Model A: (a) The readiness potential is part of the causal process leading to the *occurrences* of torn decisions, and this has nothing whatsoever to do with which option is chosen; and (b) which option is chosen is in fact TDW-undetermined. (A similar point, though a bit different, has been made by Haggard and Eimer – see, e.g., their (1999) as well as Haggard's contribution to Haggard and Libet (2001).)

Model B: (a) The readiness potential is part of the process whereby our reasons cause our decisions, and (b) in connection with torn decisions, this process doesn't determine which option is chosen; rather, it deterministically causes it to be the case that the choice will come from among the reasons-based tied-for-best options (and perhaps also that the moment-of-choice probabilities of these options being chosen are all roughly even).

Now, models A and B are both highly controversial, and as of right now, I don't think we have any good reason to endorse either of them. But the important

point here is that as of right now, we don't have any good reason to reject them either; in particular, the available evidence concerning the readiness potential doesn't give us any good reason to reject them. More generally – and in the present context, this is the really important point – as of right now, there is no reason to think that, in torn decisions, the readiness potential is part of a causal process that's relevant to the issue of which tied-for-best option is chosen. There is simply no evidence for this, and so the existence of the readiness potential doesn't give us any good reason to suppose that, in torn decisions, which tied-for-best option is chosen is causally influenced by prior-to-choice nonconscious brain processes.

Conclusion

In the last section, I responded to a variety of arguments against TDW-indeterminism. There are, of course, other arguments that one might attempt here, but I don't think any of them are cogent. In other words, at the present time, I don't think we have any good reason to reject TDW-indeterminism. And as I pointed out above, I don't think we have any good reason to *endorse* TDW-indeterminism either. Thus, if this is right, then the question of whether TDW-indeterminism is true is a wide open question. But earlier I argued that the question of whether libertarianism is true (i.e., the question of whether humans are L-free) reduces to the question of whether TDW-indeterminism is true. Thus, it seems that as of right now, the libertarian question is an open question. And in particular, it's an open *empirical* question. For (a) the question of whether we're L-free turns on the question of whether TDW-indeterminism is true, and (b) TDW-indeterminism is a straightforward empirical hypothesis about the causal histories of our torn decisions.

Cross-References

► Consciousness and Agency
► Free Will and Experimental Philosophy: An Intervention
► Mental Causation
► No Excuses: Performance Mistakes in Morality

References

Balaguer, M. (2010). *Free will as an open scientific problem*. Cambridge, MA: MIT Press.
Campbell, C. A. (1967). *In defense of free will*. London: Allen & Unwin.
Chisholm, R. (1964). Human freedom and the self. Reprinted in Watson (1982), pp. 24–35.
Clarke, R. (1993). Toward a credible agent-causal account of free will. *Nous, 27*, 191–203.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Ekstrom, L. W. (2000). *Free will: A philosophical study*. Boulder: Westview Press.

Festinger, L. (1957). *A theory of cognitive dissonance*. Palo Alto: Stanford University Press.

Fischer, J. M. (1994). *The metaphysics of free will: A study of control*. Oxford: Blackwell.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*, 829–839.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20.

Ginet, C. (1966). Might we have no choice? In Lehrer (Ed.), (pp. 87–104).

Ginet, C. (1990). *On action*. Cambridge: Cambridge University Press.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*, 128–133.

Haggard, P., & Libet, B. (2001). Conscious intention and brain activity. *Journal of Consciousness Studies, 8*, 47–63.

Hobart, R. E. (1934). Free will as involving determinism and inconceivable without It. *Mind, 43*, 1–27.

Hobbes, T. (1651). *Leviathan*. New York: Collier Books. Reprinted 1962.

Honderich, T. (Ed.). (1973). *Essays on freedom of action*. London: Routledge and Kegan Paul.

Honderich, T. (2002). Determinism as true, compatibilism and incompatibilism as false, and the real problem. In Kane (Ed.), (pp. 461–476).

Hume, D. (1748). *An inquiry concerning human understanding*. Indianapolis: Bobbs-Merrill. Reprinted 1955.

Isen, A., & Levin, P. (1972). Effect of feeling good on helping. *Journal of Personality and Social Psychology, 21*, 384–388.

Kane, R. (1996). *The significance of free will*. New York: Oxford University Press.

Kane, R. (2002). *The oxford handbook of free will*. New York: Oxford University Press.

Kornhuber, H., & Deecke, L. (1965). Hirnpotentialanderungen bei willkurbewegungen und passiven bewegungen des menschen. *Pfluegers Arch Gesamte Physiologie Menschen Tiere, 284*, 1–17.

Libet, B. (2002). Do we have free will? In Kane (Ed.), (pp. 551–564).

Libet, B., Gleason, C., Wright, E., & Pearl, D. (1983). Time of conscious intention to Act in relation to cerebral potential. *Brain, 106*, 623–642.

McKenna, M. (2012). *Conversation and responsibility*. Oxford: Oxford University Press.

Mele, A. (1995). *Autonomous agents*. New York: Oxford University Press.

Mele, A. (2009). *Effective intentions*. New York: Oxford University Press.

Milgram, S. (1969). *Obedience to authority*. New York: Harper and Row.

Nelkin, D. (2005). Freedom, responsibility, and the challenge of situationism. *Midwest Studies in Philosophy, XXIX*, 181–206.

O'Connor, T. (2000). *Persons and causes*. New York: Oxford University Press.

Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Pereboom, D. (forthcoming). Mark Balaguer's event-causal libertarianism. *Philosophical Studies*.

Reid, T. (1788). *Essays on the active powers of the human mind*. Cambridge, MA: MIT Press. Reprinted 1969.

Rowe, W. (1987). Two concepts of freedom. *Proceedings of the American Philosophical Association, 62*, 43–64.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.

Taylor, R. (1966). *Action and purpose*. Englewood Cliffs: Prentice-Hall.

Tegmark, M. (2000). The importance of quantum decoherence in brain processes. *Physical review E, 61*, 4194.

Thorp, J. (1980). *Free will*. London: Routledge & Kegan Paul.

van Inwagen, P. (1975). The incompatibility of free will and determinism. Reprinted in Watson (1982), pp. 46–58.

van Inwagen, P. (1983). *An essay on free will*. Oxford: Clarendon.

Velmans, M. (1991). Is human information processing conscious? *The Behavioral and Brain Sciences, 14*, 651–669.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

Walter, H. (2001). *Neurophilosophy of free will*. Cambridge, MA: MIT Press.

Watson, G. (1975). Free agency. *Journal of Philosophy, 72*, 205–220.

Watson, G. (1982). *Free will*. Oxford: Oxford University Press.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wiggins, D. (1973). Towards a reasonable libertarianism, In Honderich (Ed.), (pp. 31–61).

Wolf, S. (1990). *Freedom within reason*. New York: Oxford University Press.

No Excuses: Performance Mistakes in Morality

16

Santiago Amaya and John M. Doris

Contents

Abstract

Philosophical accounts of moral responsibility are standardly framed by two platitudes. According to them, blame requires the presence of a moral defect in the agent and the absence of excuses. In this chapter, this kind of approach is challenged. It is argued that (a) people sometimes violate moral norms due to performance mistakes, (b) it often appears reasonable to hold them responsible for it, and (c) their mistakes cannot be traced to their moral qualities or to the presence of excuses. In the end, the implications for discussions of moral responsibility are discussed.

Associated Press report Posted: 01/25/2013 08:18:46 AM MSTCOLONIE, N.Y. (AP) – Authorities say a New York man who left his 1-year-old son in his car for eight hours in frigid weather only realized his mistake after a call from his wife. Police in the Albany suburb of Colonie say the man forgot to drop off his son at day care and left the child strapped in the backseat of the car when he parked outside his office Thursday morning.

S. Amaya (✉)
Department of Philosophy, Universidad de los Andes, Bogotá, Colombia
e-mail: samaya@uniandes.edu.co

J.M. Doris
Philosophy-Neuroscience-Psychology Program and Philosophy Department, Washington University, St. Louis, MO, USA
e-mail: jdoris@artsci.wustl.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_120,

Introduction

Philosophical accounts of moral responsibility for wrongdoing are standardly framed by two platitudes. The first, the *positive condition*, holds that attribution of responsibility, and the associated assignment of blame, requires the *presence* in the agent of a moral defect implicated in the wrongdoing.[1] The second, the *negative condition*, requires the *absence* of excuses, such as ignorance or coercion.[2] If either condition fails to apply, the attribution of responsibility is blocked.

It is easy to see why the platitudes are platitudes. If a person's violation of a moral norm is sourced in some feature of their being that is not a moral defect – say, a mental disorder – or if a violation is committed in excusing conditions – say, under extreme duress – it is difficult to see why they should be blamed for it. Like many others, we are strongly inclined to endorse this picture. At the same time, we acknowledge the existence of a variety of *performance mistakes* in the realm of morality, where (1) people violate moral norms, (2) it appears reasonable to hold them responsible for these violations, and (3) neither of the two platitudes apply.

Something has to go. While what goes, as in other trilemmas of this kind, is a matter of contestable philosophical discretion, we will argue that neither the fact of the moral violation nor the appropriateness of responsibility attribution is readily contested. If so, it would seem that the platitudes must be discarded or revised. Given their status as platitudes, this result is of no small importance.

Our argument goes as follows. First, we introduce the notion of performance mistakes and show why they are sometimes violations of moral norms for which the actor is morally responsible. Second, we claim that these mistakes can be understood in terms of failures to exercise normative competence. With this understanding in hand, we next discuss two varieties of performance mistakes and explain why the positive and negative conditions are not there met. Finally, after having articulated the trilemma, we hint at our preferred way of addressing it.

Performance Mistakes

Performance mistakes have figured in theorizing about the human mind since the "cognitive revolution" that marked the demise of behaviorism. In his paper, "The Problem of Serial Order in Behavior," Lashley (1951) offered a disarmingly simple,

---

[1]The positive condition takes a variety of forms: Humeans require bad character (Brandt 1958; Nozick 1981), Kantians require a deficient will, (Wolff 1990; Wallace 1994), and followers of Frankfurt require motivations with origins in the "real self" (Scanlon 1998; Smith 2005). For an account of the condition that is meant to be neutral between the moral psychological pictures underlying these views, see McKenna (2012).

[2]If anything, the second platitude is even more of a platitude than the first. For examples, refer to the references given in parenthesis (Austin 1962; Wallace 1994; Baron 2007; Franklin 2011; McKenna 2012) The standard excuses are sometimes referred to as "Aristotelian conditions," after Aristotle's discussion in *Nichomachean Ethics* III.

yet extremely effective, argument against behaviorist accounts of linguistic behavior. As he noted, stimulus–response theories were utterly unable to explain a variety of performance mistakes, in particular, simple linguistic mistakes. What, Lashley asked, would be the stimulus–response explanation of the mistake in which a competent speaker reverses the order of two utterances in a spoonerism (saying "queer old Dean" instead of "dear old Queen")?

Years later, Chomsky (1959) recognized the value of this insight in his "Review of Skinner's Verbal Behavior." Lashley, in his opinion, had identified behaviorism's major blind spot: The production of grammatical utterances is not simply a matter of chained responses to stimuli, but requires implementing an abstract linguistic structure not easily read off speakers' performance. Verbal performance requires, in other words, linguistic *competence*, which Chomsky identified with tacit knowledge of a grammar, together with whatever cognitive and executive mechanisms are required for translating that general knowledge into linguistic behavior.[3]

Obviously, there are important differences between verbal and moral performances. But there are also significant parallels, as has been noted by numerous philosophers advocating "The Linguistic Analogy," the thesis that moral cognition and behavior are structurally similar to the knowledge and the deployment of a grammar in everyday speech (Harman 1999; Mallon 2008; Mikhail 2010; Roedder and Harman 2010). In the present context, the Linguistic Analogy provides a model for thinking about how moral agents can fail to behave in appropriate ways, despite having the required moral competences. Morally responsible behavior is made possible by the fact that people have a grasp of what is morally required from them and the further ability to implement that understanding in what they do. Thus, when an informed and able person makes a moral mistake, there are (at least) two possibilities. The mistake can morally impugn her motivations and sensibilities, or it can simply be a mistake in translating her moral understanding into behavior. If the later is the case, the mistake will be, like the common verbal lapses Lashley remarked upon, a performance mistake.

Moral performance mistakes should be clearly distinguished from other kinds of moral failure. For instance, to the extent that these are mistakes and not simple failures, only agents with basic moral competences can make them. So, they are unlike the failures of very young children or impaired adults, who cannot understand moral requirements or shape their behavior accordingly. Also, the moral transgressions in them are not intentional. Thus, they differ from instances of willfully malicious behavior (even from cases in which the person behaves wrongly without intent but knowingly). Lastly, to the extent that the mistakes are genuine performance mistakes, their occurrence cannot be traced to a lack of moral concern on the part of their agents; they are not instances of indifference, callousness, or the like.

Let us make this more concrete by outlining various ways in which moral performance mistakes can occur. Sometimes a thought that would normally occur

[3]Hence, Chomsky proposed a division of labor between theories of competence in linguistics and theories of performance (1965, pp. 3–4; 1964, pp. 9–10).

to you does not cross your mind; say, it simply does not occur to you to help a distressed pedestrian you see on the street (Matthews and Cannon 1975; Cohen and Spacapan 1984). Other times, the agent loses track of what she is doing; planning on dropping his child at day care on the way to work, a father winds up driving straight to the office, leaving his child in the backseat of the car all day (McLaren et al. 2005; Weingarten 2009). These and similar mistakes, it turns out, are commonplace, and sometimes have catastrophic consequences.

A central goal for theories of moral responsibility is to set criteria for responding to the behaviors of morally competent agents. At least, this is what many philosophers inspired by Peter Strawson (1962/1982) have sought to provide.[4] According to this line, when someone is appropriately attributed responsibility, the person is appropriately held to be a target of what Strawson called the *reactive attitudes*. These are a class of evaluative responses bound up with expectations about how people ought to behave toward each other, a class that includes blame-related attitudes such as indignation and forgiveness, as well as praise-related ones such as admiration and gratitude.

For the most part, responsibility theorists have focused on voluntary mistakes. If involuntary but responsible wrongdoing is discussed, the mistake is usually traced to the person's having certain disavowed biases or attitudes for which there is no good excuse. This is just an application of the *positive* and *negative* conditions for responsibility mentioned earlier.[5] Unfortunately, this schema is not easily applied to performance mistakes, despite that those who make them seem to be morally responsible for doing so and that their occurrence is normally embedded in the social practices that Strawson remarked upon.

To begin, some performance mistakes are clear cases of *wrongdoing*. The mistakes involve violations of appropriate moral norms. People are expected to help if there is a small cost in doing it, and parents are expected to not abandon their children in the car all day. At the same time, the violations are not intentional. You did not *mean* to leave your child in the car while you were at work, and the thought of *avoiding* helping did not even cross your mind. Nevertheless, you failed to display the amount of consideration or care toward someone else that was expected from someone in your position, and because of that, some of the Strawsonian *reactive attitudes* seem fitting. Resentment and indignation might seem too much. Yet, someone might reasonably be disappointed with you for not helping the pedestrian, and you would almost certainly feel guilty about having abandoned your child in the car. There may be cases where guilt is irrational, to be sure, but this certainly does not seem to be one of them: more likely there would seem to be

[4]See Vargas (2013a) for an in-depth treatment of moral responsibility from this Strawsonian perspective. Not everyone agrees with this general approach. There are, for instance, retributivist accounts that define responsibility for wrongdoing in terms of deserved suffering (Strawson 2000; Levy 2011). As it will become clear below, we contend that there are several notions in the "responsibility neighborhood," not all of which are essentially linked to punishment or reward.

[5]Examples of this trend can be found in the references given in parenthesis (Adams 1985; Scanlon 1998; Smith 2005). An exception is Sher (2006).

something amiss if, as a parent, you *failed* to feel guilt upon discovering that you left your child behind.[6]

Whether tacitly or explicitly, these attitudes seem to involve *psychological assessment*. Among the assessments that would sustain disappointment and guilt in the examples would be these: It did not occur to you to help, that you were not attending to your child, and so on. Thus, to the extent that the responses are fitting, they seem to track the fact that the actor was not in the expected state of mind. They instantiate a kind of assessment which, without being quite an evaluation of the person, still goes *beyond* the mere expression that things did not go as one would have wished (Nagel 1986, pp. 120–121; Wolff 1990, p. 41; Sher 2006, pp. 105–106). In fact, sometimes it involves putting oneself in the position of the agent and asking what kinds of considerations *ought* to have shaped one's behavior in that case.

To the extent that they are fitting, the reactive attitudes in these cases reflect the fact that in the circumstances, it was reasonable to expect the agent to behave appropriately. In this respect, they seem to track considerations of *fairness*. The mistakes do not happen because the agent was entirely lacking in the requisite capacities, or because she had only partially mastered some of the required skills, or even because their exercise required considerable effort. They happen, instead, in what one might call the agent's zone of "secure competence" (Raz 2010; Amaya in preparation). In fact, given what was required and the situation of the agent, success was not only reasonable but, as we shall see below, also *within easy reach*.

Of course, not all mistakes with significant consequences are like the performance mistakes we have been describing. Some of them *cannot* be traced to the person's being in the wrong state of mind and, thus, the psychological assessment implicit in the reactive attitudes simply cannot go through. If you trip and break a vase, plausibly, what is normally signaled by calling it a "mere accident" is that the tripping was not an indication of your state of mind at the time: You were not reckless, only clumsy. Other mistakes can be readily accounted for in terms of the *positive* condition. Their agents are not blameworthy because the relevant kind of wrongdoing is defined in terms of a specific moral defect – or, at least, a state of mind that normally indicates a defect in the agent's moral qualities – that is absent

[6]It might be objected that the attitudes that are fitting in these cases are similar, but not identical, with the reactive attitudes associated with responsibility and blame. For instance, one could argue that the fitting response in the case of child neglect is not genuine guilt but, instead, what Williams (1976) has called "agent-regret." There are complicated issues concerning the individuation of these attitudes and their relation to moral blame, which are beyond the present discussion. But if, as Williams intends, agent-regret is understood as a form of evaluation that is distinctively retrospective, agent-regret is not the attitude that fits this kind of case. *At the time* the parent had enough information to see that his behavior was inappropriate; in the light of it, what he did was *obviously* wrong. The case is, thus, unlike the examples discussed by Williams (e.g. Gaugin, the lorry driver), where the agent's decisions are regrettable or not precisely in the light of information that the agent does *not* have at the time, namely, information concerning their outcome. For discussion of the retrospective nature of agent-regret, refer to the references given in parenthesis (Williams 1976: esp. 126–128; Rosebury 1995, pp. 505–506).

in the mistake. There is, for instance, no moral blame for the clerk who gives back the wrong change simply due to an adding mistake: Theft, unlike child neglect, requires intent. Lastly, other performance mistakes are entirely excusable. Thus, the *negative* condition applies in them: A person might fail to behave appropriately because the standards were too demanding for her, or simply because the circumstances made it hard for her to live up to those standards.

Normative Competences

Philosophers who endorse the platitudes tend to agree that responsibility requires the *exercise* of certain rational competences. Put in terms of the *positive* condition, the idea is that moral defects are implicated in instances of wrongdoing, not just causally, but in the way required for responsibility and blame, by virtue of influencing what the actor finds reasonable to pursue.[7] But the point is also clear if one focuses on the *negative* condition and asks why any of the standard excuses precludes blame. Presumably, ignorance excuses to the extent that, due to lack of information, an agent cannot be reasonably expected to consider relevant reasons.

Of course, the notions of reasons and rationality are the subject of contention. But this much seems uncontroversial: A responsible agent is one that has certain kinds of normative competences (Wallace 1994, p. 226; Doris 2002, pp. 133–140), which include, at a minimum, whatever cognitive capacities are required to appreciate moral considerations, together with whatever executive capacity is required to regulate behavior in accordance with these considerations. Following standard usage in the literature, we shall henceforth refer to agents who exercise these competences as *reasons-responsive*.

Any theorist who takes the notion of reason-responsiveness as central needs to make room for the fact that people sometimes deserve blame for what would otherwise seem failures to exercise these competences. Motivated by self-interest, for instance, people sometimes disregard the interests of others; or, akratically, they give into temptation and act contrary to what they know is right. In these cases, the *positive* and *negative* conditions apply: Excessive self-interest is a defect and the presence of temptation is hardly an excuse. Thus, under the condition of reasons-responsiveness, if the agents' wrongdoing is not excusable and can be traced to a moral defect, it has to be possible to see it as an expression of which reasons they

[7]There are, of course, ways in which a moral defect can be implicated but not in the way required for responsibility and blame: if one inadvertently leaves one's child in the car on the way to do some evil work at the office. The idea that responsibility requires *exercising* certain rational competences is meant to exclude, in part, this kind of cases by defining what the *right* (i.e., non-fortuitous) relation is between wrongdoing and a moral defect. It can most perspicuously be found in so-called Reason-accounts of Kantian inspiration (Wolff 1990; Nelkin 2011). But it is also present in Frankfurtians, for whom the motivations of a real self are those whose influence in behavior the agent rationally endorses, and in Humeans, for whom the character traits implicated in responsible conduct are those that provide the agent with reasons.

are sensitive to – although not necessarily which courses of action they find most reasonable to pursue. In other words, it needs to be shown that the kinds of morally insensitive behaviors to which the platitudes apply, contrary to initial appearances, meet nonetheless the requirement of reasons-responsiveness.

John Fischer and Mark Ravizza (1998) have developed a prominent account of reasons-responsiveness designed to account for these kinds of moral insensitivity, and we shall take their account as paradigmatic here. According to it, the key for responsibility is that agents act from what they call *moderately reasons-responsive mechanisms*. The notion of mechanism is intended to include things such as conscious decision and deliberation, as well as the less reflective processes by which, say, habits and emotions shape behavior, whereas the qualification "moderately" is intended to weaken the condition in a way that allows for responsibility in cases where agents are less than ideally responsive to reasons.[8]

For Fischer and Ravizza, the requirement applies in different ways to mechanisms that realize the various cognitive and executive normative competences (1998, pp. 72–74). First, it involves the agent's exercise of a cognitive competence that they call *regular reasons-receptivity*.[9] The agent need not *recognize* all the relevant moral considerations and their relative weight, which allows blame for certain instances in which reasons are disregarded, say, due to self-interest. But her recognition of reasons must still conform to a pattern that makes sense given her values and preferences, and a general understanding of what kind of considerations count as valid moral reasons. Second, moderate reason-responsiveness involves the exercise of a kind of executive competence that they call *weak reasons-reactivity*. Motivated by competing interests, the agent may not do what she recognizes as morally reasonable. But, at the very least, if her behavior is not excusable, say, because it is being forced upon her, it has to be possible for her to act differently if she were to have enough incentives to do otherwise.

With this machinery in place, we can now outline our argument for the rest of the paper. We shall avail ourselves of the idea that responsible agency requires a moderate exercise of reasons-responsiveness to indicate why for moral performance mistakes, the *positive* and *negative* conditions fail to apply. In short, these mistakes happen when agents fail to exercise normative competences they have. Their failure to exercise them cannot be traced to moral qualities of their self.

[8]For an insightful review of the account in Fischer and Ravizza, see McKenna (2001). Among others, McKenna discusses the asymmetry between receptivity and reactivity that we discuss below, concluding that Fischer and Ravizza's requirement of weak reactivity is too weak. Here, we set aside these worries, as a strengthening of the requirement will only make our argument more straightforward.

[9]More precisely, Fischer and Ravizza require that the agent *exercise* a suitable degree of reasons-recognition, which obtains, on their view, whenever the action is brought about by a certain kind of mechanism in the agent. Thus, their account not only requires that the agent has the capacity to respond to reasons but also that such capacity be displayed to a certain degree (or "be operative" as they sometimes put it) in the workings of certain mechanisms. For an explicit statement of the view, refer to the references given in parenthesis (Fischer and Ravizza 1998, p. 53; Fischer 2006, p. 242).

And, to the extent that there were no factors preventing or making onerous their exercise, the mistakes are not excusable.

In fact, as we shall see, there are two kinds of mistakes here, corresponding to the distinction between receptivity and reactivity to reasons. First, there are *cognitive* mistakes. In them, one fails *to recognize* that certain moral considerations are germane in one's situation. But the mistake is not an indication of one's lack of acceptance of moral reasons, how they fit together, or what their relative strength is. It is rather a consequence of the cognitive routines by which, in general, social situations get structured. Second, there are *executive* mistakes. Here, one *fails to react* appropriately to reasons that one actually recognizes. But the mistake is not due to the absence of proper incentives, or to a general inability to react to those reasons. It is instead a side effect of the way in which habits and well-rehearsed routines tend to shape everyday action.

Helping Behavior

Theorists who embrace the positive condition typically emphasize moral character, reflective judgments, and personal values. Yet, there is a wealth of empirical research, now well known to philosophers, showing that moral cognition and behavior are extremely sensitive to situational factors. Arguably, this *situationist* evidence troubles traditional notions of character and personhood, but we shall not make this argument.[10] Here, we want only to illustrate a particular kind of performance mistake.

It is perhaps uncontroversial that one ought to help people in distress when there is a small cost involved in doing it. And it is perhaps equally uncontroversial that, however lenient as the requirement might be, many of us fail to live up its standards. However, the evidence suggests that these failures need not be an indication of people's deep moral convictions and motivations. Instead, there are a variety of situational factors that affect helping behavior, even on the part of people who explicitly and honestly avow values, such as compassion and care, that might reasonably be expected to promote helping. One such factor is temporal: If one is slightly late for a meeting, haste might induce one not to help someone who evidently needs it (Darley and Batson 1973). Another is social: Surrounded by an unresponsive group of people, the group's attitude might easily persuade one that help is not needed (Latané and Rodin 1969; Latané and Darley 1970).

[10]For reviews and discussion of the situationist evidence, refer to the references given in parenthesis (Ross and Nisbett 1991; Doris 2002; Miller 2003; Nahmias 2007; Merrit et al. 2010; Miller 2013; Alfano and Fairweather 2013). Elsewhere, one of us has systematically explored how this evidence raises generalized skepticism in relation to traditional notions of character and agency (Doris 2002, 2009, 2014). For discussion of how situationist evidence affects traditional accounts of responsibility, refer to the references given in parenthesis (Nelkin 2005; Vargas 2013b; Brink 2013).

It is one thing to *decide* not to help because you feel pressured about time, or because nobody else is doing it. This might show something about how you respond to moral reasons; maybe, when it comes down to it, the things that motivate you to act are not quite in synchrony with the values that you explicitly and honestly avow. And, if this is true, at least in some cases, not helping might be an expression of a moral defect, such as selfishness or callousness, giving rise to an application of the *positive* condition. Even if one is not guilty of selfishness or callousness, one might akratically forgo helping someone else, which might indicate another kind of moral defect.

It is another thing when the thought of helping *does not occur* to one, even though one typically recognizes situations where help is needed (Darley and Batson 1973, p. 107; Doris 2002, p. 138). There, it becomes harder to see how the mistake could be traced to a defect in one's character, will, or one's views about helping others. Indeed, given these things alone, it is hard to rationalize the omission. You saw an injured pedestrian in distress, you value helping others, but the question whether to help did not occur to you. *Really*?

In fact, this sort of thing seems to happen in quite ordinary circumstances. Competent moral agents often fail to recognize obvious opportunities to exhibit pro-social behaviors. And what the evidence indicates is that their lack of recognition often cannot be traced to individual traits, but it is instead a function of situational factors that do not seem to have the makings of moral reasons. That is, these are *not* factors that the agents themselves, and competent moral agents in general, would be likely to invoke in justifying their behavior, to recommend it to others, to say what was good about it, etc.

Studies of the effects of noise on interpersonal relations provide a good illustration here. In a representative experiment, Mathews and Canon (1975: study 1) found that 72 % of participants were inclined to help a confederate simply at the sight of seeing him drop a set of reading materials. However, in a situation that was exactly alike, except that the level of white noise in the room increased from 48 to 85db, which is a volume slightly higher than a telephone dial tone, the helping behavior went down to 36 % of participants.

Admittedly, in this kind of cases, it is tempting to advance explanations, seeking to trace the observed behaviors to some kind of moral defect. One, for instance, might argue that the participants in the study did consider helping, but decided against it because they wanted to escape the noise. Or one might argue that helping did not occur to them, because they are in general not very attentive to the needs of others. The problem is that, even though these explanations would help rationalize the absence of helping behavior, there is generally no additional evidence supporting them. In the study referred above, for instance, not helping did not facilitate escaping the noise – in fact, if anything, helping the confederate was the most expedient way for the participants to escape.[11]

[11](Mathews and Canon 1975, pp. 575–576). Refer to the references given in parenthesis for review of other studies (Cohen and Spacapan 1984, pp. 222–226).

More important, there seems to be an alternative explanation in the vicinity for which there is actually supporting evidence. In the presence of disturbing environmental factors, people tend to focus their attention on whatever task they are already engaged, neglecting available information that is not directly relevant for it. The phenomenon has been documented in a variety of nonsocial scenarios and abstract tasks (Cherry 1953; Broadbent 1958; Moray 1959; Wood and Cowan 1995). But there is good reason to think that it is also a factor shaping social interactions. Namely, to the extent that everyday situations do not come explicitly labeled as requiring specific prosocial behaviors, people need to exercise some kind of ethical awareness to read off what is required from them. Yet, disturbing environmental factors, such as high levels of noise, lessen attention to social cues that would otherwise indicate what the moral demands of the specific situations are (Mathews and Cannon 1975; Cohen 1978).

To test this hypothesis, Matthews and Canon conducted a further field study, in which helping behavior was again measured in low-noise (ambient noise of 50db) and high-noise (a lawn mower running at 87db) conditions (1975: Study 2). This time, they manipulated the salience of social cues: In one condition, the confederate who dropped the documents wore a cast on his right arm; in the other, he did not wear it. As before, noise level made a significant difference in observed helping. The helping cue also made a difference: Only 15 % of subjects helped the confederate with no cast, whereas 47 % helped him when he was wearing the cast. However, it only made a significant difference when ambience noise was low, which suggests that the situational factor affected behavior by pre-empting the structuring effect that social cues otherwise play.

The example illustrates the first kind of mistake we want to highlight. It is a performance mistake of a *cognitive* variety. In certain situations, due to environmental factors, people might fail to exhibit a reasonable degree of awareness and, thus, fail to see the moral dimension of the situation that lies in front of them. Their failure, however, is not indicative of their values, or a consequence of their having competing interests. They simply fail to frame the situation as one in which their moral competences are relevant. Furthermore, the failure does not seem to be evidence of a *moral defect* in their characters, or a consequence of their personal attitudes toward what is morally significant. Instead, it seems to be a function of how adult human beings generally allocate attention and frame the situations they face in the presence of disturbing environmental factors.

Recall Fischer and Ravizza's notion of regular *reasons-recognition*. According to them, being adequately responsive is a matter of acting on mechanisms that exhibit an understandable pattern of recognition of reasons, at the very least, one in which the reasons actually and potentially recognized track one's values and weighted preferences. In the kind of mistakes we are now considering, however, this evidently is not the case. The reasons recognized at the time depend upon how one interprets the situation, but the mechanisms that shape the interpretation are influenced by factors, such as ambient noise, that are of little practical significance. Their influence is not a reflection of one's values and preferences; they are not even among the kind of things that have the makings of moral reasons (Doris 2005, p. 657, 2009, p. 66).

Now, in so far as these mistakes involve a failure to exercise some of the rational competences required for moral responsibility, it would seem that the *positive* condition does not obtain: The mistakes do not show much, or anything at all, about one's moral qualities as a person. On the other hand, the *negative* condition does not seem to apply either. To be sure, the presence of situational factors might excuse in the sense of being a mitigating consideration. Their presence prevents tracing the mistake to a moral defect in the agent. Yet, it does not seem that the standard excuses work here, as implied by the *negative* condition, by making it unfair to hold the person responsible.

To see this, begin with *duress*. Intuitively, the presence of ambient noise is not coercive in the way in which having a gun pointed to one's head is. Yet, in so far as noise affects helping behavior in a significant number of cases, one might be inclined to conclude that the pressure is intense enough to make the corresponding moral violations excusable. The problem is that statistical considerations like this are hardly ever, at least not by themselves, sufficient to indicate that behaving appropriately posed an onerous demand. Overstaying a parking meter, for example, is common, but the fact that it happens often gives no reason to suppose that the infractions occur under duress.

There are undoubtedly situational pressures that are so intense they result in genuine cognitive impairment. The circumstances of war, for example, can seriously degrade one's cognitive competences (Doris and Murphy 2007). However, the mistakes we are discussing here are not like this. In them, there seems to be a clear disproportion between the modest situational input and the insensitivity to moral considerations exhibited in the mistake – just to compare, whereas in the Mathews and Cannon study, noise level in high-noise conditions went up to 85db, a single gunshot may produce a sound over 160db. And it is this disproportion, at least in part, which seems to preclude the mistake in the cases considered from being excusable due to impairment.[12]

Of course, situational factors can also induce forms of temporary impairment that fall short of cognitive degradation. Despite her best efforts, a competent student might choke under pressure and, thus, be excused for bombing an exam in which she would otherwise perform aptly (Baumeister 1984; Beilock 2007). But, without any further evidence, to model the effect of noise discussed here along these lines would seem too radical. Levels of ambience noise in New York City's subway platforms, for instance, routinely cross the 85db level. Yet one would be hard-pressed to conclude on these grounds that many of the four million weekday subway riders are temporarily impaired.[13]

[12]For discussion of the extent to which situational factors, in general, can actually impede pro-social behavior, refer to the references given in parenthesis (Doris 2002, pp. 128–153; Nelkin 2005, pp. 193–194; Vargas 2013; Brink 2013).

[13]For data of noise level in NYC transportation system, refer to the references given in parenthesis (Neizel et al. 2009). Díaz and Pedrero (2006) measured sound exposure in different locations in Madrid, several of which presented averaged continuous sound levels over 85db: nightclub, soccer pitch, musical theater, primary school (exit door), and carwash.

Consider, on the other hand, *ignorance*. Certainly, there might be cases where external factors make a situation ambiguous and, thus, difficult to know whether one is required to help or not. Group effects, for instance, are perhaps explained by the fact that one takes the non-responsive attitude of other people as evidence that one's help is not needed. But clearly not all cases are ambiguous. In Matthews and Canon second study, for example, the pedestrian in distress was evidently alone, had a shoulder-to-wrist cast, and was visibly unable to collect by himself the books scattered on the ground.

Obviously, any situation in which a person behaves poorly by not realizing what is morally required at the time involves some kind of ignorance. In so far as the question whether to help never occurred to you, it is true in some sense that you did not know that you had to help there. But treating this kind of ignorance as exculpatory would be a stretch, for it does not involve the unavailability of information, or the inability to gather it. Indeed, at the time all the information necessary to make the right call was available to you and, as a competent moral agent, you could and should have inferred from this information that your help was needed.

Death by Hyperthermia

We now turn to a different kind of moral performance mistake that concerns what we earlier referred to as reasons-reactivity. The literature on human performance and human factors engineering contains a wealth of examples suggesting that many of the mistakes people make are not a reflection of deep seated attitudes in them but are rather due to small lapses of concentration and memory. Interestingly, these *slips* do not happen randomly but come in systematic patterns, which suggests that they are not isolated glitches but rather mistakes in the life of altogether normal human agents (Norman 1981; Reason and Mycielska 1982; Amaya 2013, Amaya in preparation).

Every summer in the USA, there are on average 30 reported cases of children who die of hyperthermia after being inadvertently left in the backseat of a car (McLaren et al. 2005; Weingarten 2009). In comparison to the top three causes of unintentional death of children in the country – motor vehicle accidents, drowning, and burns – the rate of deaths by hyperthermia pales. Yet, given that in at least 19 out of 50 states there are laws regarding unattended children in vehicles, 556 deaths since 1998 is by all means a grim figure.[14]

[14]Sher (2009) discusses a hypothetical case, "hot dog," that looks like the performance mistake exemplified by these hyperthermia cases. However, as he constructs his example (and other cases he discusses), the mistake is meant to illustrate a different kind of failure—one in which the mistake reflects psychological states constitutive of the agent's individual character. Even if Sher's interpretation of his hypothetical case is accepted, it is not clear that it applies generally to the real cases discussed here.

How is this *remotely* possible? Perhaps, some parents simply do not care enough about their children, or care too much about their busy lives. But, as always, if one looks at the specifics, things are less clear. Sure, some adults are simply thoughtless: They use the backseat of their car as substitutes for day cares. Some of them have deep problems managing their lives. In one instance, a drunk parent drove home, went into the house, passed out, and left the child in the car overnight. These, however, are just some of the cases: Devoted and competent parents are also touched by this tragedy (Collins 2006; Weingarten 2009). And even though they seem responsible for the death of their children, their mistakes do not seem traceable to a moral defect in them and, hence, a candidate for the *positive* condition.

Rather, a common thread behind the stories involves a change in habitual routines. During the week, you and your partner take turns dropping the child at the day care. One a given morning, however, you switch the usual schedule – your partner has an early meeting – so you seat the child in your car before leaving home. Minutes later, however, instead of taking the freeway exit for the daycare, you head straight to work. Your child is quiet. Children tend to fall asleep in a moving car.

Slips similar to these are common in everyday life. In a now classic study, Reason (1984) asked participants to keep a diary where they would record the slips they made during 1 week, indicating what they intended to do at the time, what they wound up doing, the circumstances in which it happened. At the end of the week, Reason had an inventory of 192 slips, a list with items perhaps too prosaic to deserve much attention, except that many of them shared a similar structure with genuinely catastrophic slips: "I meant to get my wallet from the bedroom. Instead, I wound the bedside clock and came down again without the wallet." As in the parental slip, in these cases, the person sets to act with an intention to do something (e.g., fetch the wallet). Yet, in the course of doing it, she gets distracted, a well-rehearsed routine (e.g., winding the bedside clock) kicks in, and she winds up derailed midway.

Diary studies raise a host of concerns; people have considerable difficulty in accurately reporting on themselves. But the general trend noticed above is one that has been observed in subsequent diary studies and in many "accident" reports collected for insurance and legal purposes.[15] When people fail to monitor their performance adequately, their behavior will automatically tend to follow well-rehearsed routines consistent with their immediate surroundings and their general know-how. As a consequence, what they wind up doing is not necessarily in line with what they intended to do, or the reasons that first motivated them to act (Reason 1990, p. 184; Amaya in preparation).

Thus, a slip constitues a further kind of performance mistake. It is in this case an *executive* kind of mistake. Here the person recognizes what she needs to do. The parent, say, recognizes that the child needs to be taken care of, carefully buckles her in the safety seat in the back, and cautiously drives out to the day care. The problem

[15]For a discussion of possible pitfalls in diary studies and self-reports of slips, refer to the references given in parenthesis (Morris 1984; Reason 1993). For a review of more recent studies and discussion of Reason's original results, refer to the references given in parenthesis (Sellen 1990; Jónsdóttir et al. 2007).

is failing to follow through, to execute the full plan of action. It is a matter of the agents not reacting appropriately to reasons involving the well-being of the child that she already has recognized as germane.

Think back to Fischer and Ravizza's notion of *reasons-reactivity*. According to them, in addition to the recognition of reasons, reasons-responsiveness requires *weak reactivity*. It requires, at least, that one acts on a mechanism that would function differently if the right sort of incentives were in place. Thus stated, the requirement is nicely met in the akratic case: If you were offered a $1,000 dollars to not eat the cake and stick to your diet, you will likely forgo the temptation. The case of the parental slip, however, is different. In it, all the incentives that would otherwise motivate the parent to act differently are already in place; it is not as though the parent values getting his workday off to an early start more than the life of his child, and it is unlikely that increasing the penalty for the mistake would help prevent it by increasing the cost of ending one's child's life.[16] What happens, instead, is that he acts on a habitual routine and habits tend to become somewhat rigid, which means that they need not track how moral considerations align in specific situations. That is, they are mechanisms that can guide behavior without reacting to the moral reasons that the agent accepts.

It bears repeating what we previously said regarding situational influences. In some sense, saying that the mistake is a slip functions as an excuse: It serves to signal lack of bad will, absence of intent, etc. But this is not quite yet to exculpate. For accepting that the mistake is not in the purview of the *positive* condition does not remove it altogether from the purview of practices associated with the attribution of moral responsibility. The *negative* condition, in other words, does not apply either: It is still a case of wrongdoing where some blame-related reactive attitudes seem fitting.

Again, think about the standard excuses. To begin, observe that invoking *duress* here would not work. Slips, we know, are more prevalent in situations of extreme stress, for example, an emergency in an airplane cockpit (Broadbent et al. 1986; Reason 1988). Yet, even though parenting a young child could be stressful, it does not seem that parents who make these slips are under significantly more stress than the many parents to which the tragedy never happens. Stress might mitigate; it does not exculpate.

It would seem that *ignorance* is a bit more promising. Obviously, it would not be normative ignorance: The parent knows what good parenting requires. Nor, strictly speaking, would it be factual ignorance, for the parent knows all the facts necessary to see the wrongness of his behavior. But perhaps, there still is room for some kind of epistemic mistake that makes the slip excusable. The parent, after all, does not

[16]In fact, even though in the US system, judges are generally moved by the parents suffering to give lenient sentences, there are a significant number of cases in which parents receive extremely harsh punishments: 15 years to life prison sentences or spending one day in jail on the child's birthday for 7 years. For an empirical study of the treatment of hyperthermia cases in the US judicial system from 1998 to 2003, see Collins (2006).

knowingly leave the child in the back of the car (if he knew it, he would not have done it).

The problem here is that it is not obvious what kind of epistemic mistake this could be. Notice, first, that false belief is not a good candidate. True, the parent's omission would seem more reasonable, say, if at some point, he would have come to believe falsely that the child was not in the car, or that he had already dropped her at the day care. However, given the explanation of the slip provided above, attributing the false belief to the parent to explain the mistake seems redundant (and perhaps an ad hoc move motivated only to get the *negative* condition right). The parent gets distracted, a habitual routine kicks in, and he winds up driving straight to work. The question that the putative false belief is supposed to answer, namely, what were the whereabouts of the child, never even comes up.[17]

To make this clear, distinguish mistakes that happen because what one knows fails to shape one's behavior from those in which one omits doing something due to the belief that that thing no longer needs to be done. Once the distinction is in place, the slip looks more like a case of the former, not the latter kind. The parent's belief that the child is in the back of the car at some point becomes inactive: The parent ceases to attend to it. But this is not to say that the parent comes to form a false belief about where the child is.

Now, it is true that this kind of epistemic failure sometimes excuses. Plausibly, one could be excused for not fulfilling a promise made years ago, if one temporarily forgets it. And one might also be excused if, despite trying really hard, being in a tip-of-the-tongue state, one cannot recall the name of the person with whom one is conversing. Notice, however, that the slip is unlike either of these cases in which one is excused for not acting in accordance with what one knows. The parent does not make the mistake because he fails to recall an event of a distant past, or because he tries without success to remember where his child is. In fact, it was well within his competence to bring to mind whatever he needed to have present at the time. Before the tragic day, he effortlessly did it several mornings a week.

Finally, it is worth asking to what extent the slips considered so far are *accidents*. Here, the point is a bit less straightforward. For, on the one hand, being the result of performance mistakes, slips are not committed on purpose and fall somewhat short of being instances of self-governance. Yet, on the other hand, the agent seems to have sufficient control of the actions that take place when the slip happens. The parent's drive is not a tick or stumble. He drives to work the way he does it on any given day in which his partner (not him) is on daycare duty.

To appreciate why considerations of accidentality might not be an issue here, contrast slips mistakes with paradigmatic episodes of sheer bad luck. Part of what makes the phenomenon of moral luck so hard to accommodate is that the distinction between, say, the unlucky driver who hits a pedestrian and the lucky driver who

[17]Unfortunately, in some cases, the slip does result in the person having a false belief. Having spent all day at work, some parents drive to the day care in the afternoon to pick up their child, not having realized their slip. But as an explanation of how the mistake first came about, attributing the false belief clearly gets the order of things wrong.

does not is not one that is traceable to the state of mind of the drivers (Nagel 1976; Williams 1976). But in the present case, the state of mind makes all the difference.[18] You knew all you needed to know in order to prevent the tragedy. But you left the child behind, because you were not keeping track of where she was.

This is not to say, of course, especially with catastrophic slips, that there is not something unlucky about them. It is surely a sad set of coincidences that children fall asleep in moving cars, that cars warm up so quickly, and that passers-by tend not to look at the interior of parked cars. But, in the end, it is not clear that such factors, which are admittedly beyond one's control, make the relevant difference. Would you cease to be responsible for leaving your child behind if some stranger were to break into your car and rescue her?

Conclusion

Performance mistakes do not result from a moral defect in their agents. And many of them do not occur under circumstances that make them excusable. Yet, there seem to be good reasons to hold them responsible for them. These kinds of cases are, therefore, a challenge to accounts of responsibility framed in terms of the *positive* and *negative* conditions stated at the beginning of this chapter. That is, they seem to challenge picture of responsibility and blame circumscribed by agentic moral qualities and excusability.

Again, it looks like something has to go. We might (1) argue that moral performance mistakes meet the positive and negative platitudes, or (2) jettison the positive and negative platitudes from our account of responsibility. Neither option is appetizing. Given the centrality of the platitudes in traditional and contemporary theories of moral responsibility, (2) is radically revisionary. And given what we know about the psychological mechanisms underlying the performance mistakes described here, (1) seems implausible.

There is, however, another solution, one that requires neither giving up the appearances nor giving up the platitudes. According to it, performance mistakes should be viewed, not as straightforward counter-examples to existing accounts that

[18]Here we refer to the kind of luck that is normally described as resultant luck (Nagel 1979; Nelkin 2013). Some philosophical accounts of luck, formulated in terms of chanciness, might seem applicable here: Roughly, a chancy event is one that fails to occur in a large enough number of possible worlds obtainable by making a small change in the actual world (Pritchard 2005; Coffman 2007, p. 390). Even assuming the correctness of this kind of approach to common instances of luck (see Lackey 2008 for discussion), it is questionable whether chanciness *thus understood* excuses. An agent might introduce small changes in counterfactual scenarios and that may very well mean the difference between the agent being or not being blameworthy. The parent, for instance, could put his briefcase in the backseat, which will "remind" him that his child is still in the car before anything happens to her. Of course, things might be different in relation to other kinds or accounts of luck. Levy (2011), for instance, argues that it is a matter of (constitutive) luck whether and what kind of moral considerations cross your mind at the time of any action. But notice that he advances this claim to defend his general skepticism about human beings ever being responsible.

emphasize moral qualities and excusability, but as evidence that there are several notions in the moral responsibility "neighborhood" and that existing theories capture only some of them. Performance mistakes would, thus, fall under the ambit of a notion that, while in "the responsibility family," does not require the positive and negative platitudes be met for its application.[19]

The distinction, perhaps, can be made in terms of answerability vs. accountability – but beware that these terms has been used slightly differently by others (Watson 1996; Shoemaker 2011). Answerability, on the proposed view, would mark the positive relation between a person's behavior and her qualities as a moral agent. Accountability, on the other hand, would mark the relation that the person bears to those performance mistakes illustrated here, where the violations of moral expectations are brought about by a combination of the circumstances of her action and the dynamics of the cognitive and executive capacities than agents normally rely on.

From the perspective adopted here, accountability without answerability would be a form of moral responsibility. Even though it is not dependent on agentic appraisal, it goes further than a mere negative assessment of behavior. It reflects the fact that competent moral agents are bound by reasonable expectations; that, in the absence of excusing conditions, violations of these expectations are disapproved; and that the disapproval is fittingly expressed in attitudes of blame and guilt. The difference would be that, because the violations are not traceable to anything morally deficient in their agents, recovering from the mistakes does *not* require sanction or punishment.[20]

Cross-References

▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
▶ Consciousness and Agency
▶ Drug Addiction and Criminal Responsibility
▶ Free Will, Agency, and the Cultural, Reflexive Brain
▶ Moral Cognition: Introduction
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play

References

Adams, R. (1985). Involuntary sins. *Philosophical Review, 94*, 3–31.
Alfano, M., & Fairweather, A. (2013). Situationism and virtue theory. *Oxford Bibliographies in Philosophy*.

[19]For a pluralistic approach to moral responsibility, refer to the references given in parenthesis (Doris et al 2007; Doris 2014).

[20]We would like to thank Neil Levy and Michael McKenna for insightful comments, as well as audiences at the University of Lund, University of Texas, El Paso, and the Berlin School of Mind and Brain for helpful criticisms and suggestions.

Amaya, S. (2013). Slips. *Noûs, 47*(3), 559–576.
Amaya, S. (in preparation). *Without belief*. Manuscript in preparation.
Austin, J. L. (1962). A plea for excuses. In J. O. Urmson & G. J. Warnock (Eds.), *Austin, philosophical papers* (3rd ed.). Oxford: Oxford University Press.
Baron, M. (2007). Excuses, excuses. *Criminal Law and Philosophy, 1*, 21–39.
Baumeister, R. (1984). Choking under pressure: selfconsciousness and paradoxical effects of incentiveson skillful performance. *Journal of Personality and Social Psychology, 46*, pp. 610–620.
Beilock, S. (2007). Choking under pressure. In R. Baumesiter & K. Vohs. (Eds.), *Encyclopedia of Social Psychology*, Thousand Oaks, CA: Sage Publications, pp. 140–141.
Brand, R. (1958). Blameworthiness and obligation. In A. I. Melden (Ed.), *Essays in moral philosophy* (pp. 3–39). Seattle: University of Washington Press.
Brink. D. (2013). Situationism, responsibility, and fair opportunity. *Social Philosophy and Policy*.
Broadbent, D. E. (1958). *Perception and communication*. New York: Pergamon.
Broadbent, D. E., Broadbent, M., & Jones, L. (1986). Correlates of cognitive failure. *British Journal of Clinical Psychology, 21*, 1–16.
Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25*, 975–979.
Chomsky, N. (1959). Review of B.F. Skinner's verbal behavior. *Language, 35*(l), 26–58.
Chomsky, N. (1964). *Current issues in linguistic theory*. The Hague: Mouton.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Coffman, E. J. (2007). Thinking about luck. *Synthese, 158*, 385–398.
Cohen, S. (1978). Environmental load and the allocation of attention. In A. Baum, J. Singer, & S. Vain (Eds.), *Advances in environmental psychology*. Hillsdale: Erlbaum.
Cohen, S., & Spacapan, S. (1984). The social psychology of noise. In D. M. Jones & A. J. Chapman (Eds.), *Noise and society* (pp. 221–245). New York: Wiley.
Collins, J. (2006). Crime and parenthood: The uneasy case of prosecuting negligent parents. *Northwestern University School of Law, 100*, 807–856.
Darley, J. M., & Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*, 100–108.
Díaz, C., & Pedrero, A. (2006). Sound exposure during daily activities. *Applied Acoustics, 67*, 271–283.
Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. New York: Cambridge University Press.
Doris, J. M. (2005). Replies: Evidence and sensibility. *Philosophy and Phenomenological Research, 71*, 656–677.
Doris, J. M. (2009). Skepticism about persons. *Philosophical Issues, 19*, 57–91.
Doris, J. M. (2014). *Talking to ourselves*. New York: Oxford University Press.
Doris, J. M., Knobe, J., & Woolfolk, R. (2007). Variantism about responsibility. *Philosophical Perspectives, 27*, 183–214.
Doris, J. M., & Murphy, D. (2007). From my lai to Abu Grahib: the moral psychology of atrocity. *Midwest Studies in Philosophy, 31*(1), 25–55.
Fischer, H. (2006). *My way: Essays on moral responsibility*. Oxford: Oxford University Press.
Fischer, J., & Ravizza, M. (1998). *Responsibility and control*. Cambridge: Cambridge University Press.
Franklin, C. (2011). A theory of the normative force of pleas. *Philosophical Studies*. doi:10.1007/s11098-011-9826-y.
Harman, G. (1999). Moral philosophy and linguistics. In K. Brinkman (Ed.), *Proceedings of the 20th world congress of philosophy* (pp. 107–115). Bowling Green: Philosophy Documentation Center.
Jónsdóttir, M., Adólfsdóttir, S., Cortez, R. D., Gunnarsdóttir, M., & Gústafsdóttir, H. (2007). A diary study of action slips in healthy individuals. *The Clinical Neuropsychologist, 21*, 875–883.

Lackey, J. (2008). What luck is not. *Australasian Journal of Philosophy, 86*, 255–267.

Lashley, K. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior. The Hixon symposium* (pp. 112–136). New York: Wiley.

Latané, B., & Darley, J. (1970). *The Unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.

Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander information. *Journal of Experimental Social Psychology, 5*, 189–202.

Levy, N. (2011). *Hard luck: How luck undermines free will and responsibility*. New York: Oxford University Press.

Mallon, R. (2008). Reviving Rawls's linguistic analogy inside and out. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 2, pp. 145–156). Cambridge, MA: MIT Press.

Matthews, K., & Cannon, L. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology, 32*, 571–577.

McKenna, M. (2001). Review of "Responsibility and control: A theory of moral responsibility". *Journal of Philosophy, 98*, 93–110.

McKenna, M. (2012). *Conversation and responsibility*. New York: Oxford University Press.

McLaren, C., Null, J., & Quinn, J. (2005). Heat stress from enclosed vehicles: Moderate ambient temperatures cause significant temperature rise in enclosed vehicles. *Pediatrics, 116*, e109–e112.

Merrit, M., Doris, J. M., & Harman, G. (2010). Character. In J. M. Doris & The Moral Psychology Groups (Eds.), *Moral psychology handbook* (pp. 354–400). New York: Oxford University Press.

Mikhail, J. (2010). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge: Cambridge University press.

Miller, C. (2003). Social psychology and virtue ethics. *The Journal of Ethics, 7*, 365–392.

Miller, C. (2013). *Moral character: An empirical theory*. New York: Oxford University Press.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology, 11*, 56–60.

Morris, P. E. (1984). The validity of subjective reports. In J. Harris & P. Morris (Eds.), *Everyday memory, actions, and absent-mindedness* (pp. 173–190). New York: Academic Press.

Nagel, T. (1976). *Proceedings of the Aristotelian Society, 50*, 137–151.

Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press.

Nahmias, E. (2007). Autonomous agency and social psychology. In M. Maraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the mind* (pp. 169–185). Dordrecht: Springer.

Neizel, R., Gershon, R., Zeltser, M., Canton, A., & Akram, M. (2009). Noise levels associated with New York City's mass transit systems. *American Journal of Public Health, 99*, 1393–1399.

Nelkin, D. (2005). Freedom, responsibility, and the challenge of situationism. *Midwest Studies in Philosophy, 29*, 181–206.

Nelkin, D. (2011). Making sense of freedom and responsibility. Oxford: Oxford University Press.

Nelkin, D. (2013). Moral luck. In E. N. Zalta (Ed.). Retrieved from *The Stanford encyclopedia of philosophy* (Summer 2013 Edition). http://plato.stanford.edu/archives/sum2013/entries/moral-luck/

Norman, D. (1981). Categorization of action slips. *Psychological Review, 88*, 1–15.

Nozick, R. (1981). *Philosophical explorations*. Cambridge, MA: Harvard University Press.

Pritchard, D. (2005). *Epistemic luck*. Oxford: Oxford University Press.

Raz, J. (2010). Being in the world. *Ratio, 23*, 433–452.

Reason, J. (1984). Lapses of attention in everyday life. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 515–549). New York: Academic Press.

Reason, J. (1988). Stress and cognitive failure. In S. Fisher & J. Reason (Eds.), *Handbook of life stress, cognition, and health* (pp. 405–421). New York: Wiley.

Reason, J. (1990). *Human error*. Cambridge: Cambridge University Press.

Reason, J. (1993). Self-report questionnaires in cognitive psychology: Have they delivered the goods? In A. Baddeley & L. Weiskrantz (Eds.), *Attention: selection, awareness, and control* (pp. 406–424). Oxford: Clarendon Press.

Reason, J., & Mycielska, K. (1982). *Absent-mindedness?* Englewood Cliffs: Prentice-Hall.

Roedder, E., & Harman, G. (2010). Linguistics and moral theory. In J. M. Doris (Ed.), *The moral psychology handbook* (pp. 273–296). New York: Oxford University Press.

Rosebury, B. (1995). Moral responsibility and "moral luck". *The Philosophical Review, 104*, 499–524.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. Philadelphia: Temple University Press.

Scanlon, T. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.

Sellen, A. J. (1990). *Mechanisms of human error and human error detection*. Ph.D. thesis, University of California, San Diego.

Sher, G. (2006). *In praise of blame*. New York: Oxford University Press.

Sher, G. (2009). *Who knew? Responsibility without awareness*. New York: Oxford University Press.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics, 121*, 602–632.

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental Life. *Ethics, 115*(2), 236–271.

Strawson, P. F. (1962/1982). Freedom and resentment. In G. Watson (Ed.), *Free will*. New York: Oxford University Press.

Strawson, G. (2000). The unhelpfulness of determinism. *Philosophy and Phenomenological Research, 60*, 149–156.

Vargas, M. (2013a). *Building better beings: A theory of moral responsibility*. Oxford: Oxford University Press.

Vargas, M. (2013b). Situationism and responsibility: Free will in fragments. In T. Vierkant, J. Kiverstein, & A. Clark (Eds.), *Decomposing the will* (pp. 400–416). New York: Oxford University Press.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*, 227–248.

Weingarten, G. (2009). Fatal distraction. *The Washington Post*, March 08, 2009. http://www.washingtonpost.com/wpdyn/content/article/2009/02/27/AR2009022701549.html. Retrieved 1 Feb 2010.

Williams, B. (1976). Moral Luck. *Proceedings of the Aristotelian Society, 50*, 115–135.

Wolff, S. (1990). *Freedom within reason*. New York: Oxford University Press.

Wood, N. L., & Cowan, N. (1995). The cocktail party phenomenon re- visited: Attention and memory in the classic selective listening procedure of Cherry (1953). *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 255–260.

Free Will and Experimental Philosophy: An Intervention

17

Tamler Sommers

Contents

Abstract

This chapter reviews and then criticizes the dominant approach that experimental philosophers have adopted in their studies on free will and moral responsibility. Section "Experimental Philosophy and Free Will" reviews the experimental literature and the shared approach: probing for intuitions about the so-called compatibility question, whether free will is compatible with causal determinism. Section "The Intervention" argues that this experimental focus on the compatibility question is fundamentally misguided. The critique develops in the form of a dialogue: a staged "intervention" for an experimental philosopher who works on free will. The chapter concludes with some thoughts about how the literature can move in a more fruitful direction.

Introduction

Though not an experimental philosopher myself, I have always been a supporter of the movement and have had little patience the hostility that has too often been directed its way. The most common and tiresome objection to experimental

T. Sommers
University of Houston, Houston, TX, USA
e-mail: tssommers@uh.edu; tamlers@gmail.com

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_118,

philosophy – I do not care about people's *intuitions* about [problem X], I am searching for the *truth* about X" – is especially misguided when applied to the problem of free will and moral responsibility, since virtually all philosophical theories on this topic – both compatibilist and incompatibilist – rely on appeals to intuition. The arguments in these theories typically operate under the assumption that we have certain intuitions about crucial cases or premises and go on to draw substantive philosophical conclusions on that basis. Experimental methods can shed light on (a) the degree to which these intuitions are shared by members of a larger community, and (b) the psychological mechanisms underlying that underlie the intuitions. Thus, experimental philosophy can help us understand the "truth" about free will and moral responsibility in ways that complement more traditional "armchair" analysis.

The general question, then, of whether experimental philosophy can make important contributions to this topic has an obvious answer. Yes, it can.[1] But has it? In recent years, I have become quite critical of the particular approach that experimental philosophers have adopted in their investigations.[2] My fear is that the approach has become entrenched – "industry standard" as one experimental philosopher has put it – which makes it harder to abandon in spite of its serious flaws. Even worse, these flaws play into the hands of the movement's critics. If my analysis is correct, experimental philosophers are testing for intuitions that the critics are *correct* not to "care" about, because those intuitions do not shed light on the truth of existing philosophical theories. As someone who sees enormous potential in its tools and methods, I find this deeply troubling. So rather than join the chorus of X-phi bashers, I have decided to do what all friends should when someone close to them engages in self-destructive behavior: stage an intervention. This intervention, in the form of a dialogue, will take place in section "The Intervention". Before that, I will summarize some foundational studies and then outline the shared approach that I find so problematic.

Experimental Philosophy and Free Will

For reasons not clearly understood by anyone, philosophers have focused most of their attention on the so-called compatibility question, the question of whether free will and moral responsibility are compatible with the truth of causal determinism. Unsurprisingly (but disappointingly), experimentalists have followed suit. Beginning with Nahmias, Morris, Nadelhoffer, and Turner's (2006) seminal studies, philosophers have attempted to directly probe folk intuitions on the compatibility question. Nahmias and colleagues presented participants with a series of

[1]See my "In Memoriam: The X-Phi Debate" on why there is no reasonable debate over this general question.

[2]See Sommers (2010, 2012) for two examples.

vignettes in which agents perform an act of wrongdoing in a deterministic world. Participants were then asked whether the agents acted of their own free will and whether they are blameworthy for the act. In each of these scenarios, a majority of the participants gave compatibilist answers to both questions. In one scenario, for example, an agent robs a bank in a determined world and 76 % of the participants judged the agent acted of his own free will and 83 % responded that he was morally blameworthy. Nahmias and colleagues received complementary results for the praiseworthy and morally neutral scenarios as well. According to the authors, these results cast doubt on the claims of philosophers, such as Robert Kane and Galen Strawson, that ordinary people – referred to (somewhat unfortunately) as "the folk" in X-Phi literature – are "natural incompatibilists."

In their influential paper, Nichols and Knobe (2007) use a similar approach to explain the asymmetry between incompatibilist claims about folk intuitions and the experimental results in the studies by Nahmias et al. They hypothesize that in philosophy seminars, when considering the compatibility question abstractly, we tend to have incompatibilist intuitions. But in concrete cases that trigger emotional responses, our intuitions become more compatibilist. To test this hypothesis, Nichols and Knobe describe two universes to their participants, a deterministic universe (Universe A) in which everything including human decision-making is completely caused by events tracing all the way back to the beginning of the universe, and the other (Universe B) where everything *with the exception* of human decisions is completely caused by past events. The key difference, according to the scenarios:

> is that in Universe A every decision is completely caused by what happened before the decision—given the past, each decision *has to happen* the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision *does **not** have to happen* the way that it does. (p. 669)

Participants were then divided into a concrete high affect condition and an abstract low affect condition. In the low affect condition, participants were simply asked if people could be fully morally responsible for their actions in this deterministic universe. Here, a large majority (86 %) of the participants answered "no." The high affect condition describes a man named Bill who burns down his house, killing his wife and three children, so that he can be with his secretary. Participants are then asked if Bill is fully morally responsible for his behavior. In this condition, 72 % of the participants answer "yes." Nichols and Knobe offer these results as an explanation for the intractability of the free will problem. When we consider the problem abstractly, one set of cognitive processes lead us to the conclusion that determinism is incompatible with free and responsible action. But cases like Bill trigger a different set of processes that dispose us to assign blame and responsibility for terrible crimes and worry less about how they were caused.

A related factor that may affect our intuitions on freedom and responsibility is *psychological distance*: the distance (either in space or time) between participants and the event or object and events they are considering. Weigel (2011) asked participants to imagine hearing a lecture about a deterministic universe, and

were then asked if a murderer in this universe acted freely. Some participants were assigned to a condition in which the lecture on determinism was taking place in a few days. Others were assigned to a condition in which the lecture took place in a few years. This seemingly small manipulation had a significant effect. The results showed that participants were less inclined to say that this man freely decided to kill his family when they imagined hearing about it at a more distant time. Research on psychological distance suggests that greater distance triggers cognitive processes that deal with questions more abstractly, so these results lend support to the view that our conflicting intuitions on free will are the product of different cognitive processes.[3]

Feltz and Cokely (2009) also adopt this general framework in their studies, but add an interesting twist: The authors investigate whether personality differences can affect intuitions on free will. Specifically, Feltz and Cokely predicted that participants who were high in personality trait extroversion would be more likely to attribute free will and moral responsibility to a murderer in the deterministic scenario. Rather remarkably, the results showed a significant correlation between extroversion and a willingness to attribute free will and responsibility for determined behavior. These results suggest yet another reason for why philosophical reflection and debate alone have not led to more universal agreement about the free will problem.

In the above studies, participants are asked to imagine a world where human behavior is caused deterministically, but they are given no detail about the nature of the causation. Several studies have shown, however, that the *type* of causal explanation can influence free will and responsibility judgments. Monterosso et al. (2005) found that participants were more likely to exonerate agents from responsibility for crimes when the explanation for their action was physiological (e.g., a chemical imbalance) in nature rather than experiential (e.g., abusive parents). Nahmias, Coates, and Kvaran (2007) and Nahmias and Murray (2010) separated participants into two conditions. In the first, agents' decision-making is described "in terms of neuroscientific, mechanistic processes"; in the second, decision-making is described "in terms of psychological, intentional processes ("psych scenarios")". They found that participants in both abstract and concrete cases regarded the neuroscientific descriptions of decision-making as more threatening to free will and responsibility than psychological descriptions.

[3]An experiment by Roskies and Nichols (2008), although not designed explicitly to test the effect of psychological distance, supplies further evidence of its influence The authors predicted that intuitions about free will and moral responsibility would be sensitive to whether deterministic scenarios are described as *actual*, in our world, or merely *possible* (true in some other possible world). In the "alternate condition," the description of determinism is the same, except that the universe the subjects are asked to consider is not our own. The subjects then respond to similar probes about free will and moral responsibility in this alternate universe. Consistent with the authors' hypothesis, the assignments of free will and moral responsibility were significantly higher in the actual condition than in the alternate condition.

Nahmias and colleagues offer these results as evidence for an “error theory” for incompatibilist intuitions. The results, they argue, show that participants are prone to confuse determinism with *fatalism*, the view that our conscious desires and deliberations does not causally influence our behavior and destiny. But this is an error – determinism does not entail that our conscious deliberations do not cause our behavior, only that the deliberations are themselves determined by previous causal factors. The authors conclude that folk intuitions might be more unified in favor of the compatibilist view that we can be free and responsible as long as our actions are determined in the right way, i.e., through our deliberation.

Most of the other experimental work on free will and moral responsibility has employed this same basic approach, which can be outlined as follows.

1. **Select the variable to be manipulated.** Examples described above are concrete versus abstract cases, affect-laden versus affect- neutral cases, psychological distance, temperament, and type of causation. Other studies have manipulated the culture of the participant (Sarkissian et al. 2010), examined patients with emotional deficits (Cova et al. 2012), and the world in which the wrong-doing in the vignette takes place (our world or a different one) (Roskies and Nichols (2008)).
2. **Describe a world in which all human action is determined.**[4] With a few important exceptions, most studies have adopted the description provide by Nichols and Knobe (2007)[5]. Participants are asked to imagine that determinism is true, or asked hypothetical questions about what would be true in such a world.
3. **Probe for intuitions about free will and moral responsibility.** Finally, the participants are asked whether agents can be free or morally responsible for their behavior in this world. The behavior differs depending on the study and the manipulation, but typically the author will describe an example of wrongdoing in the determined world and then ask whether the agents acted of their free will and can justly be held morally responsible for this action.

With several notable exceptions, this is the general approach that experimental philosophers have adopted in their investigations into this topic.[6] As noted above, I believe this approach has several flaws that are serious enough to make an intervention necessary.

[4]Monterrosso et al. (2005), to their credit, do not discuss determinism. Instead, they vary the percentages of people with a particular condition—psychological or physiological—who go on to commit crimes.

[5]Eddy Nahmias objects that the “had to happen” language in this description begs the question against compatibilist analyses of choice. His studies do not employ this description.

[6]See Sripada (2012) and Feltz (2012) for examples of studies that stray from this approach, and thus are not vulnerable to the central criticism of this chapter.

The Intervention

The following is a transcript of a conversation with an Experimental Philosopher Who Works on Free Will and Moral Responsibility (acronym: ETHAN). Three friends of the experimental philosopher—Emma, Sarah, and Maynard—*have asked Ethan to come for a brief "healing" session to get him to address some recurring problem with his work.*

Emma: The first thing I want to make absolutely clear is that we're here to help you. Everyone here is in your corner. We're not here to be dismissive, to tell you that your projects are not "really philosophy" or that they have no value.

Maynard: Not that you deserve all the fawning media attention you get either, but. . . .

Sarah: Stop that. We're here to help Ethan, not lash out.

Ethan: What exactly are all of you talking about?

Emma: We all got together and agreed that we needed to have, well, a conversation with you. We think you've been in a bit of a self-destructive spiral recently. We know you can pull yourself out of it, but first you have to recognize what's happening. We don't want things to get to a point where you've lost all everything and you're just handing out flyers in Washington Square park like a. . .

Ethan: Isn't that how this experimental philosophy thing got started?

Emma: Ok, bad example. We don't want him descending to a point where he's camped out in the woods on the outskirts of some university just. . .

Sarah: Again. . .

Emma: Fine! Let's just get this over with, ok?

Ethan: My sentiments exactly.

Sarah: Let's begin with the elephant in the room, the biggest problem of all, the one that's plagued your work from the beginning, the days of Nahmias, Nadelhoffer, Morris, and Turner. A lot of the other problems stem from this one. Now Ethan—what is the whole justification for probing the "folks" about their intuitions?

Ethan: The folk.

Sarah: Whatever. Why is probing for intuitions important in the first place?

Ethan: You know the answer to that. Philosophers appeal to intuitions all the time, that's how they defend their theories of free will. They appeal to intuitions in their arguments. I say this at the beginning of practically every paper I've published.

Emma: And rightly so, good. But you can't just probe for any intuitions you want, right? Which intuitions would you want to probe?

Ethan: Intuitions about free will and responsibility. Obviously.

Maynard: So you just want their random fragmented thoughts about free will?

Ethan: You know what I mean. The intuitions that are at center of the debate.

Sarah: Right. And even more specifically, wouldn't you want to test for the same intuitions that the philosophers are appealing to in their theories and arguments?

Ethan: Of course, that's obvious.

Maynard: If it's so obvious why aren't you doing it?

Ethan: What do you mean? I am doing it.

Maynard: No you're not. Let me ask you something—how many, say, incompatibilist arguments are there that go like this?

1. If determinism is true, then intuitively there is no such thing as free will.
2. Therefore, determinism is incompatible with free will.

Does that sound like a good argument to you? Or would that be a just little question-begging?

Ethan: Well, that would be question-begging. Although Robert Kane and Galen Strawson do say that people start out as naturalists incompatibilists and have to be talked out of it by the clever arguments of compatibilists.

Maynard: We know, trust us. You quote the same damn passages in all your papers. And you know where every one of those passages come from? Introductions. They aren't part of any theory, not part of any argument. They add nothing of substance. They're throwaway lines. *They play no role in incompatibilist arguments*, they're just there to set the mood. So let me ask you again: is that what you think is so important to test? Intuitions about introductory throwaway lines in incompatibilist books and articles?

Ethan: I'm not sure I fully understand....

Emma: What he's trying to say is that you're testing intuitions on the wrong question. You're probing intuitions on the compatibility question itself. You're testing intuitions about the *conclusions* of compatibilist and incompatibilist arguments. But philosophers don't appeal to intuitions about conclusions. They appeal to intuitions about the *premises* of their arguments, intuitions about cases or narrower principles like transfer principles or PAP principles.[7]

Ethan: Ok, right, I see. You think that people's intuitions about those principles or cases might be different than their intuitions about the compatibility question itself.

Sarah: Exactly.

Ethan: That's an interesting hypothesis, you should test that.

Maynard: Here's an idea. How about after someone makes an objection to a study or a whole series of studies, you don't just sidestep it by telling them they should test it? You know what *real* scientists never say when people raise objections to their studies? "That's an interesting hypothesis, you should test that." You want to know why they don't say that? Because they know that other people have their own lives, their own labs, their own careers, and they're not going to drop everything just because of the problems with another psychologist's experiment.

[7]The transfer of non-responsibility principle, roughly speaking, is that we cannot be morally responsible for an act if we are not morally responsible for any of the determining factors of that act. (The non-responsibility for the determining factors "transfers" to the act itself). See Fischer and Ravizza (1998) for a more precise formulation. The Principle of Alternate Possibilities (PAP) is the principle that an agent can only be morally responsible for an action if they had the capacity to act otherwise.

Ethan: But Josh Knobe says that all the time.

Sarah: Look, Ethan. I know this is difficult. But Joshua Knobe, much as we all admire him and love him, important and influential as he is, Josh Knobe is not God.

Ethan: I know that.

Maynard: Say it then. Josh Knobe is not God.

Ethan: I just said it.

Emma: (gently) No you didn't.

Ethan: (fidgeting uncomfortably) Ok, fine. He's not.

Maynard: Not what?

Ethan: (more fidgeting). Not God.

Maynard: Who's not God.

Ethan: Josh Knobe, OK? Josh Knobe is not God. Satisfied?

Maynard: The point is that this a glaring flaw with *your* experiments, *your* whole way of approaching this topic. *We* are working on other things have other commitments.

Emma: This is a serious problem, Ethan. You're not asking the right question. And it's plaguing so much of your work these days.

Sarah: It really is—at the very least, four out of every five X-phi studies use this model. There is so much creativity and potential in experimental philosophy, so many fascinating ideas, but we have no idea what to make of the results because of the flaws of the model. Look at this study that's just coming out by Florian Cova and colleagues (2012): "Judgments about Moral Responsibility and Determinism in Patients with Behavioural Variant of Frontotemporal Dementia: Still Compatibilists." What a cool idea, right? Testing patients with emotional deficits. We can learn so much—there are so many potential implications for Strawsonian approaches to moral responsibility to name just one. But what is he doing in the study? He's giving these patients the same questions that Nichols and Knobe gave in their original *Nous* article—questions that test the *conclusions* of arguments. And based on their responses he concludes that the patients are "still compatibilists." But you can't say that they're still compatibilists unless you ask the right question! You're wasting so many golden opportunities.

Ethan: Look, these questions are industry standard now.

Emma (to Sarah): Avoidance.

Maynard: Are you listening to us? That's our whole point: the industry standard is seriously screwed. And anyway, Nichols and Knobe came out in 2007 for Christ sake. The Nahmias et al. papers came out in 2006. How much of a standard can there be in a few years?

Emma: You need to change the industry standard, or else every other future paper will have this exact same problem that undermines everything you're trying to do.

Ethan: Ok then how about the recent Chandra Sripada (2012) article that probes for intuitions about manipulation cases and then for intuitions about whether manipulated agents are acting from their 'deep self.'

Emma: Right, good! That's exactly the kind of approach that you need to adopt—because it's testing *premises* of real philosophical arguments! But that paper is the exception. It strays from the "industry standard" and it should.

Sarah: Let me point out something else. The original classic X-phi study—the famous Josh Knobe in Washington Square park experiment with the CEO and the environment—that worked precisely because it was testing for the right intuitions (See Knobe 2006). Intuitions about a particular case. No determinism, no ambiguous words. Did the CEO intend to help or hurt the environment? That's it. Now that's a model of the kind of thing you need to be doing.

Emma: And this leads us to the second problem we think you need to face up to.

Ethan: Christ, what now?

Emma: These concepts that you're asking about in the studies are far too complex and ambiguous for non-philosophers. They can be interpreted in so many different ways. This means there's too noise in your experiments even if you were probing for the right intuitions.

Ethan: Can you be just a little more specific?

Sarah: Well, I hate to pick on Nichols and Knobe, but they can handle it. Let's go back to their study. They give their participants a description of a determined universe, we talked about the problems with that already. But then they ask their participants: Can an agent, or can Bill, be "fully morally responsible" for his actions? Now, think about that: "fully morally responsible," what is that supposed to–

Maynard: I've worked in this field my whole career and I have no clue what the hell that means.

Emma: The point is that 'fully morally responsible' can be interpreted in a lot of different ways. I mean: is anyone ever *fully* morally responsible? External factors play *some* role in every action, right? Nobody disagrees with that.

Ethan: So how would you put it then?

Sarah: Good question. To be honest, every time I teach this topic in an intro course it takes me at least 15 or 20 minutes to nail down the sense of moral responsibility that's at the center of the debate.

Ethan: What sense is that?

Sarah: Moral desert, deserving blame or punishment independent of the consequences.

Ethan: Fine but you can't say 'moral desert.'

Emma: That's right, you can't. That's a meaningless term to a 19 year old non-philosophy major. And even philosophers disagree about how to analyze moral desert.

Sarah: And when I bring up responsibility in my class, we've already discussed non-consequentialism so I can use terms and concepts that wouldn't work for your surveys of non-philosophers. That's why I can get the concept across in only 15 or 20 minutes. But "non-consequentialist desert" is certainly not what most people think of when they hear "moral responsibility" if they think about it at all.

Emma: It's a really tough concept, right? But as philosophers who work on this issue, *we* know what we mean by morally responsibility (for the most part), so we've forgotten all the different ways that people interpret it and how confusing it can be.

Ethan: Ok, so what about 'blameworthy' and 'praiseworthy'?

Emma: I don't think so. Those are still philosopher's terms. People don't use that language outside of a class or maybe a courtroom.

Maynard: Nichols and Knobe is not even the worst example. How about Roskies and Nichols asking if people "*should be morally blamed* for their actions" in a determined universe. That's confusing on so many levels.

Ethan: Now hold on. That paper was published in the *Journal of Philosophy*!

Maynard: Oh my god you're right! It must be flawless then. Can you stop avoiding the issue? First of all, what do they mean by 'morally blamed'—have you ever heard anyone use that term in your life? Even worse, someone could think people should be blamed for consequentialist reasons even if they don't deserve it.

Sarah: And this is our point. It's one thing to avoid getting bogged down by trivial philosophical distinctions. We're all in favor of that. We're not fastidious word-choosers, we don't get aroused by conceptual analysis. But it's another thing to play so fast and loose with terms that you have no idea what your participants mean by their judgments. That's exactly the kind of noise you have to minimize in a study.

Ethan: Let me tell you something: it's pretty easy to just sit there in your armchair and criticize our terminology and our work. But I haven't heard you suggest any better options.

Maynard: Don't play the martyr with us. It's not our job to write your studies for you. The problem with you X-Phi people is that you get too much unfair criticism, all those hostile strawman objections from philosophers who feel threatened by experimental philosophy. So now you have a bunker mentality. You can't see that there might be legitimate serious problems out there.

Sarah: Ethan, we understand that it's tough getting that language right. I've thought about this a lot and I have no idea how to phrase those questions. It's a serious difficulty. Maybe you need to get a little more clever and oblique. Maybe you can do some behavioral studies where someone gets punished at a cost for no other reason than they deserve it.

Maynard: And keep it as close to real life as possible. For the love of God, no more planet Erta scenarios or hypothetical worlds where alien scientists have developed supercomputers to detect neuronal activity and predict everything that will happen. Who the hell knows what are intuitions mean when they're completely divorced from reality.

Ethan: Philosophers use wacked out thought experiments all the time! You can't blame *that* on me.

Maynard: Philosophers hold their big job meeting between Christmas and New Years in cold weather cities. Does that make it right? Moral responsibility is a *human* phenomenon. Our intuitions respond to actual things that happen to us, not to ludicrous science fiction fairy tales. We probably learn something from probing judgments about alien worlds and whacked out fantasies about evil neurosurgeons, but nothing resembling what we want to learn.

Ethan: I actually agree with that. But look, this is a new field. We're learning on the fly. How do you think that we're going to make any progress if we don't try new things?

Emma: I'm so glad to hear you say that, Ethan. What it shows is that you're starting to recognize the flaws in your approach. You're making progress.

Ethan: Great, thanks. If only you didn't sound so patronizing. . .

Maynard: You want less patronizing? I can handle that. I have a couple things to get off my chest and I promise you it won't sound patronizing.

Sarah: Let's stick to the plan, OK? The next item on the agenda is how Ethan is starting to overreach a bit. Overstating the significance of certain results. . .

Maynard: Overreach?? Are you freaking kidding me? I want to read something to you. You wonder why you have so many hostile unsympathetic critics, right? Well here's some prime ammunition for them. All of your problems distilled in a couple pages. This is from Knobe and Nichols in the new edition of the Kane free will anthology. Let me read this to you. They're discussing a possible objection, one suggesting that different cultures might possibly have different intuitions about free will and responsibility than some college students at Florida State or wherever. And that certain intuitions in the debate may be shaped by cultural forces like the American emphasis on individual autonomy. You want to talk about patronizing? Here's how they respond:

> We certainly agree that these [diversity claims] are very plausible hypotheses but the empirical evidence thus far has not been kind to them. In a recent study, subjects from India, Hong Kong, Colombia and the United States were all presented with the abstract condition of the experiment described above (Sarkissian, Chatterjee, De Brigard, Knobe, Nichols & Sirker forthcoming). Strikingly, the majority of subjects in all four of these cultures said that no one could be fully morally responsible in a deterministic universe, and there were no significant cross-cultural differences in people's responses. Yet ordinary people, many of whom have never thought about these questions before, seem somehow to immediately converge on one particular answer. In fact, we find this convergence even across four different cultures, with radically different religious and philosophical traditions. What could possibly explain this striking pattern of intuitions? (Knobe and Nichols 2011, p. 532)

Ethan: Right, that's an interesting result. What's the problem?

Maynard: Well, let's break this down. "The empirical evidence has thus far not been kind" to the diversity hypothesis. What does this unkind empirical evidence amount to? One study. Is there a mass of empirical evidence that suggests otherwise? Yes. Is any of it discussed? No. But OK, that evidence doesn't specifically concern the compatibility question (which is itself an obsession of philosophy in the West, but just set that aside for now). So maybe they're after something more specific. All right, let's look at the study. They give the Nichols and Knobe deterministic scenario to university students in Hong Kong, India, and Columbia. University students! They supposedly represent the essence of every culture. They then ask is it's possible for someone to be morally responsible for their behavior in that deterministic universe. We already talked about the fundamental flaw in this approach—you're testing intuitions about the conclusion of all the compatibilist and incompatibilist arguments. But set that aside too for now. Pretend that's not a problem. They're still basing this incredible conclusion—universal convergence on an ancient philosophical

problem—on *one study*. One study that gets a null result! Let me repeat that: the null result of a single study. That's the evidence that "hasn't been kind" to the diversity hypothesis. A null result—who the hell knows why you didn't get significant differences? There are so many possibilities. Could something be wrong, I don't know, with the methodology? Could that be why you didn't get significant differences? And we've talked about the problem with 'fully morally responsible' in English. What word are you using in Chinese? As far as I know, there isn't even Chinese word for moral responsibility in the sense that we're talking about.

Ethan: The Hong Kong students were fluent in English. They did the study in English.

Maynard: Great, even better! These Hong Kong students who are fluent in English are supposed represent ordinary people in a different culture with "with radically different religious and philosophical traditions." And they're supposed to interpret "fully morally responsible" just like the Americans, and just like me I guess, even though I still don't have a clue what that means. And of course, all the participants in each country are (a) university students, (b) from high SES backgrounds, and (c) of university age. These are your "striking," "deeply puzzling and mysterious" results that cry out for explanation. A null result from the judgments of university students in a second language in a study with ambiguous terminology that asks the wrong question. Do you see the problem? Do you see why people get mad at experimental philosophers?

Ethan: You seem to be taking this study personally.

Emma: I think he knows someone who wrote a book on the diversity of intuitions about moral responsibility.

Ethan: Who?

Maynard: Nobody, that's not the point, let's stick to the issue.

Sarah: We don't have problems with the study so much—well, besides the ones we've already talked about. The issue here is more with the wild claims that Knobe and Nichols are making about it. The study is, at best, an intriguing first step. It might even be suggestive if it weren't for the problem that it tests the wrong intuitions. But you have to admit: it's completely implausible to conclude that "ordinary people, many of whom have never thought about these questions before, seem somehow to immediately converge on one particular answer."

Maynard: Let's put it this way. You're always complaining about armchair philosophers who pull their claims about people's intuitions out of their you know what. But is what you're doing here any better?

Ethan: At least we're trying to test for intuitions in these other countries. That's something, isn't it?

Maynard: It's something, but it might be worse in the end than what the armchair people are doing because it's bathing itself in the haloed pretense of real empirical support.

Ethan: (Breaks down, sobbing). You're right. Oh God I see it now! I'm so so soooorrry.

Emma: Good, good... let it out.

Sarah: It's Ok, we're here for you.

Maynard: This is bullshit, right?

Ethan: (stops sobbing immediately). Yes, I figured that's what I was supposed to do. But seriously, I get the point. Can we wrap this up?

Emma: I don't know, can we? What have we learned from this so far?

Ethan: Look, if you're waiting for me to really break down and start sobbing, that's not going to happen.

Sarah: We don't expect that, we just want an acknowledgment that you understand that you have some problems and that you intend to address them. Again, we're your friends and you have our full support.

Ethan: Ok, yes, I'll try to address them.

Maynard: Address what?

Ethan: All right, fine. First, I have to stop giving scenarios with deterministic worlds, and then asking if people can be morally responsible in them.

Emma: Because?

Ethan: Because when I do that, I'm not probing for the intuitions that philosophers are appealing to in their arguments. I have to start testing for intuitions about cases and principles, and investigating the source of those intuitions.

Sarah: Exactly!

Ethan: Second, I need to nail down the sense of moral responsibility we're after. Not that you had any good ideas about the best way to do that but... that's my responsibility.

Maynard: In a different sense of the word, yes it is.

Ethan: Ha ha, you should be headlining in the Catskills. Third. I need to avoid outlandish thought experiments and bring my scenarios down to earth as much as possible. Fourth, I need to stop exaggerating the implications of the data from my studies, even though most social psychologists and cognitive scientists do that all the time.

Emma: Fair enough. But that doesn't make it right.

Ethan: No it doesn't. And fifth–

Sarah: Let me just say, Ethan, that although we've been a little harsh with you today, we are doing it out of love and respect. You've brought so much energy and excitement to our field, breathed new life into it. And we think you have the potential to significantly improve our understanding of free will and moral responsibility.

Emma: Amen. We all agree with that, right Maynard?

Maynard: (grumbling) Right. Yes.

Sarah: Continue Ethan.

Ethan: Thank you. Fifth—and most important—the next time my friends invite me to a so-called "intervention" about my work, I need to tell them to get a life and mind their own business.

Cross-References

- ▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
- ▶ Determinism and Its Relevance to the Free-Will Question
- ▶ Free Will, Agency, and the Cultural, Reflexive Brain
- ▶ Neuroscience, Free Will, and Responsibility: The Current State of Play

References

Cova, F., et al. (2012). Judgments about moral responsibility and determinism in patients with behavioural variant of frontotemporal dementia: Still compatibilists. *Consciousness and Cognition, 21*(2), 851–864.
Feltz, A. (2012). Pereboom and premises: Asking the right questions in the experimental philosophy of free will. *Consciousness and Cognition, 22*, 53–63.
Feltz, A., & Cokely, T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition, 18*(1), 342–350.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 13*(2), 203–231.
Knobe, J., & Nichols, S. (2011). Free will and the bounds of the self. In R. Kane (Ed.), *Oxford handbook of free will* (pp. 530–554). Oxford: Oxford University Press.
Monterosso, J., Royzman, E., & Schwartz, B. (2005). Explaining away responsibility: Effects of scientific explanation on perceived culpability. *Ethics and Behavior, 15*(2), 139–158.
Nahmias, E., & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New waves in philosophy of action*. New York: Palgrave-Macmillan.
Nahmias, E. Morris. S., Nadelhoffer, T., & Turner, J. (2006). *"Is Incompatibilism Intuitive?" Philosophy and Phenomenological Research 73*(1), 28–53.
Nahmias, E., Justin Coates, D., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy, 31*(1), 214–242.
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663–685. Print.
Roskies, A. L., & Nichols, S. (2008). Bringing moral responsibility down to earth. *Journal of Philosophy, 105*(7), 371–388.
Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language, 25*(3), 346–358.
Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass, 5*(2), 199–212.
Sommers, T. (2012) *Relative Justice*. Princeton University Press.
Sripada, C. S. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research, 85*(3), 563–593.
Weigel, C. (2011). Distance, anger, freedom: An account of the role of abstraction in compatibilist and incompatibilist intuitions. *Philosophical Psychology, 24*(6), 803–823.

Section IV

Neuroanthropology

Toward a Neuroanthropology of Ethics: Introduction

18

Juan F. Domínguez D.

Contents

Abstract

The main argument of this introductory chapter is that there is a pressing need for a neuroanthropology of ethics because the neural bases of moral agency are to be found beyond the confines of a single brain: in the coming together and interacting of a community of brains, in the shaping of the moral brain by the social field and culture, and in the workings of a neurocognitive system that evolved to absorb, reproduce, and contribute to shared worlds of meaning. This chapter shows how the papers in this section demonstrate the need for a neuroanthropology of ethics that is sensitive to context, culture, history, and diversity as well as to the relationship between universalism and particularism, scientific fact and personal experience. This chapter also outlines areas of future research in neuroethics from a neuroanthropological perspective. A neuroanthropology of ethics is well placed to unsettle long-held assumptions about ethical behavior. It also offers new ways of approaching ethical phenomena and opens up exciting new avenues of enquiry.

J.F. Domínguez D.
Experimental Neuropsychology Research Unit, School of Psychological Sciences, Monash University, Melbourne, VIC, Australia
e-mail: juan.dominguez@monash.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_21,

Anthropology and Ethics

Understanding different ethical systems and diverse forms of moral life has been at the core of anthropological enquiry. "Ethics" and "morality" have often been indistinguishable from "culture," "ideology," or "discourse," concepts used in anthropology to represent shared rules, ideas, values, beliefs, opinions, desires, and preferences. This is not surprising as society was equated with morality for a long time in anthropology, ever since Durkheim defined society as a system of moral facts (Laidlaw 2002; Zigon 2008). As a result, anthropologists have documented in their ethnographies of societies the world over the set of ideas, values, preferences, and practices that underpin local ethics. Adding to this rich history, over the past 15 years there has been renewed anthropological interest in ethics as a separate analytical field. The methodological and ontological ground of this new anthropology of ethics is the individual, but this is a relational individual who has undergone substantial socialization. An ethical subject, from this perspective, is "always of intersubjective, social and cultural tissue" (Faubion 2011, p. 120). This recentering of the object of study has allowed anthropological enquiry to move beyond questions of social regularity and control where individual moral behavior is understood as acquiescence or transgression (Laidlaw 2002; Zigon 2008). The anthropology of ethics has in particular started to consider issues having to do with how and where ethical subjectivities form; with reference to which particular set of ideas, values, preferences, and practices are these subjectivities formed; how are they expressed in ordinary, everyday social encounters; to what extent does ethical behavior manifest through the adherence to or creative transcendence of normative action; what is the role of reflexivity and habitus in the expression of ethical behavior; and to what extent individuals inhabit different ethical subjectivities (Laidlaw 2002; Zigon 2008; Lambek 2010).

A distinctively anthropological perspective has therefore much to contribute to neuroethics. An anthropological neuroethics can first and foremost draw upon anthropology's virtually inexhaustible trove of ethnographic data to test the universality of neuroethical claims; to investigate the local conditions that modulate brain responses associated with the formation of ethical values, behaviors, and predispositions; or to better understand the relationship between context and the neural correlates of ethical behaviors. An anthropological neuroethics can further draw upon anthropology's increasingly sophisticated understanding of ethical subjects (as intersubjective, as the product of socialization and enculturation, as inhabiting various ethical positions, as reproducing and transcending normative action, etc.) to guide enquiry into the neural bases of moral agency (see also ▶ Chap. 9, "The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering"). But at a more fundamental level, an anthropological neuroethics is necessary for the same reason that an anthropological neuroscience is ultimately necessary: because human individuals are not pure individuals; because they cannot be more fully understood either in isolation or as an abstraction; and because they are cultural individuals, intersubjective individuals, individuals who inhabit (and contribute to) a world of meaning shared with others. The neural bases of moral agency, like the neural bases of human activity more broadly, cannot therefore be understood without reference to this anthropological individual.

Neuroanthropology and Ethics

The aforementioned anthropological neuroscience, or neuroanthropology, as a field of study interested in "the experiential and neurobiological aspects of cultural activity" (Domínguez et al. 2010, p. 140) is already in existence. In a great display of foresight, the field was first formulated over three decades ago [by a small group of scholars including Charles Laughlin (D'Aquili et al. 1979), contributor to this section], well before new and powerful technologies (including not only brain imaging but also high performance scientific computing) and analytic approaches brought about the current neuroscientific revolution. Recently, interest in neuroanthropology has been bolstered as a result of these technical innovations and by the picture of the human brain they are starting to draw, which is in better accord with evolving anthropological models of human behavior. Such a brain is plastic, plural, and highly responsive to the environments it is embedded in, including "the human body, the social field, and the longer-term, larger-scale structures that are the product of human brains yoked into cooperative [and competitive, it should be added], cumulative [but also ever changing] systems" (Lende and Downey 2012, p. 2), that is, cultural systems. A striking demonstration of the brain's responsiveness to these environments is that they account for a large proportion of the brain's structural variability, as measured by volume variation across the brain (Brun et al. 2009); in other words, these environments literally shape the brain (particularly white matter, the wiring of the brain that allows different brain areas to communicate, and the frontal cortex, which is the seat of higher cognitive functions, including executive control and reasoning). Moreover, cultural neuroscience, a field of study that has overlapping aims with neuroanthropology, has provided extensive evidence of cultural influences on brain function, all the way from perceptual judgments (Hedden et al. 2008), to attentional control (Ketay et al. 2009), emotional responses (Chiao et al. 2008), theory of mind (Kobayashi et al. 2007, 2008), self-knowledge, and self-construal (Zhu et al. 2007; Han et al. 2008; Lewis et al. 2008; Chiao et al. 2010).

In addition to cultural influences on brain structure and function, neuroanthropology is interested in questions such as how are socially shared meanings and practices represented in the brain; how does the brain appropriate and act on cultural experiences; what are the neural mechanisms that make culture possible; how did these mechanisms evolve; and how do they interact with genetically mediated neurocognitive processes and behaviors (Domínguez et al. 2010). Examples of neuroanthropological research addressing these questions include functional neuroimaging of role playing (Whitehead et al. 2009); neural encoding of basic types of social relationships (Iacoboni et al. 2004); neural substrates of the disposition to treat someone as a rational agent (Gallagher et al. 2002); sociocultural patterning of neural activity during self-reflection (Ma et al. 2012); social interaction research through simultaneous two-person brain imaging or hyperscanning (Konvalinka and Roepstorff 2012); relational coding as the principle of prefrontal cortex function as well as the basis for the sharing of experience (Domínguez et al. 2009); and differences in resting-state activity between human and chimpanzee brains (Rilling et al. 2007).

An anthropological neuroethics is clearly necessary because, as the above account implies, the neural bases of moral agency are to be found beyond the confines of a single brain, in the coming together and interacting of a community of brains; in the shaping of the moral brain by the social field and by those collective long-term, large-scale systems of ideas, values, models, preferences, and practices known as cultures; and in the workings of a neurocognitive system that evolved to absorb, reproduce, and contribute to shared worlds of meaning (Domínguez et al. 2009; Whitehead 2012).

Another aspect of neuroanthropological research of importance for neuroethics is neuroanthropology's critical, reflexive, and meta-analytic undertakings, which speak not only to the practice of neuroscience but also to the implications of neuroscientific knowledge for people's understanding of their own behavior, and to political and social uses of this knowledge (see also ▶ Chap. 33, "Historical and Ethical Perspectives of Modern Neuroimaging", and ▶ Chap. 40, " Neuroimaging Neuroethics: Introduction") (Domínguez et al. 2010; Domínguez 2012; Lende and Downey 2012). Researchers have, for example, problematized the process that leads to the generation of brain images as scientific "facts" by drawing attention to how experimental design, data gathering strategies, analysis, and interpretation of results incorporate and reinforce cultural assumptions about human nature, normality, brain function, and the relationship between experience and the brain (Dumit 2004); highlighting how components of this process are obscure to most researchers who must accept that subjects are transformed into objective brain images as if by magic, by automated, largely black boxed computerized procedures (Roepstorff 2002); and bringing to light the discursive concealing and marginalization of subjects and subjectivity in brain imaging resulting in the scientific process and its products gaining an inflated appearance of validity (Domínguez 2012). In addition, Turner (2012) has called into question the ontological status of the terminology used to describe mental and cognitive constructs and has recommended that a scientifically more useful ontology may result from ethnographic reports of such constructs across cultures. Dumit (2004) has also critically written on the social history of brain images after they leave the laboratory detailing how these images, rather than *illustrative* of "textual and quantitative proof" (Dumit 2004, p. 142) as intended in scientific papers, have come to be seen in popular magazines (see also ▶ Chap. 92, "Neuroscience, Neuroethics, and the Media"), in the eyes of policy makers, or the courtroom (see also ▶ Chap. 84, "Neuroimaging and Criminal Law"), as the *proof itself* of a normal brain, a depressed brain, or an insane brain; and as straightforward, objective, unmediated photographs of these categories. Lende and Downey (2012) have, on their part, underscored the need for neuroanthropology to study how "laboratories, government institutions, funding bodies, and professional organizations" fashion neuroscience knowledge and practice, to then "examine how the biomedical industry as well as lay people draw on this knowledge to understand and manage issues related to health and well-being" (see also ▶ Chap. 69, "Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues") (Lende and Downey 2012, p. 7).

Contributions in This Section

The aim of the foregoing discussion has been to show that anthropology is doubly relevant for neuroethics: not only has anthropology a solid and evolving corpus of knowledge regarding ethics and morality but also a growing body of understandings having to do with the relationship between culture and the brain. The ingredients are therefore there for a neuroanthropology of ethics. The papers in this section are a first formal contribution in this direction. In the opening paper, "The sense of justice: a neuroanthropological account," Charles Laughlin critiques the highly abstract, elaborate, hyperational, and ethnocentric view Western philosophers have of justice and proposes instead what he calls a *sense of justice* [justice, sense of] in order to better capture an attribute that intuitively seems to be shared by people everywhere. According to Laughlin, the sense of justice is inherently relational and emerges from linking a basic intuition for fairness or balance with a capacity for empathy and the experience of positive and negative affect. Laughlin argues that these component elements of the sense of justice are mediated by discrete neurophysiological mechanisms, which have precursors in other big-brained social animals. For Laughlin the sense of justice is universal, but it manifests differently depending on social structural and cultural factors. He offers examples of how the sense of justice plays out in specific contexts using ethnographic cases (including from his own fieldwork) and by showing that the sense of justice fulfills different functions in egalitarian and less stratified band and tribal societies compared to hierarchical, highly differentiated, bureaucratized societies: it is a structuring principle of social institutions among the former but becomes alienated from social procedures and juridical institutions in the latter.

In "Free will, agency, and the cultural, reflexive brain," Steve Reyna starts, like Laughlin, with a critique of Western philosophical conceptions about the will, concerned, as they have been, with a functional account of this construct but neglecting altogether its underlying substance, its structure, its materiality, and the set of mechanisms that give rise to it. Reyna identifies the brain as the structure of will and the brain's reflexivity as a key mechanism giving rise to will. By reflexivity Reyna means the brain's capacity to monitor the outer and inner milieus. Reflexivity thus understood engenders will as it constitutes a resource for finding out what is the state of the world to then figure out what to do about it. However, in the case of the human brain, action is biased by culture, which constitutes a set of understandings about the state of the world and what to do about it that are *socially shared*. The corollary is that "a culturally reflexive brain performs the functions of will." Finally, Reyna adopts a critical stance by arguing that free will (see also ▶ Chap. 13, "Neuroscience, Free Will, and Responsibility: The Current State of Play," ▶ Chap. 14, "Consciousness and Agency," ▶ Chap. 17, "Free Will and Experimental Philosophy: An Intervention," and ▶ Chap. 15, "Determinism and Its Relevance to the Free-Will Question"), understood as "unrestricted action," is inconsistent with a brain that has a definite structure defined by biology or culture or both. For Reyna, acts of will are a consequence of the brain's biological and cultural determinants and biases. He further argues that the notion of free will is often deployed as

a tool of domination by those in power, who may allocate responsibility for acts of transgression (like theft) entirely on the "free will" of those transgressing without reference to social structures (e.g., of poverty) with a causative role. Instead, Reyna advocates the use of the alternative concept of "agency" as it better incorporates such structures and explicitly articulates issues of power in the expression of will.

In the last contribution to this section, "What is normal? A historical survey and a neuroanthropological perspective," Paul Mason argues against an objective basis for the concept of normality, tracking down its historical roots, following its development, highlighting its inconsistencies, dissecting its uses and their contexts, and denouncing its abuses. Mason exposes normality as a tool of homogenization and essentialization, as a tool for obscuring the diversity of human phenomena and, as such, as an instrument of control in the hand of powerful, interested actors. Diversity is, according to Mason, not merely obscured but often made out to be degenerate. The concept of degeneracy, Mason shows, has itself been discursively maligned as part of the normalizing drive, its neutral meaning of divergence from a type being co-opted by a morality of deviance and decay. Mason reviews the effect of normality on the neurosciences where diversity in brain structure and function is reduced by transferring individual brains to a standard space; by using standardized, average results as markers of kinds of people; by generalizing the results of brain research focused on a very narrow band of the human population (largely young, university students of contemporary Western industrialized nations); and by reducing to neurobiological explanation complex problems that are in reality product of heterogeneous "intersecting variables that take place along the multistranded life course of each individual." In light of this, Mason recommends neuroscience to embrace a view of diversity mediated by an alternative, morally neutral, conception of degeneracy whereby variability is a condition of complex systems rather than a sign they are breaking apart (see also ▶ Chap. 111, "Neuroethics of Neurodiversity").

Future Directions

The contributions in this section draw from the vast and well established but ever developing body of anthropological understandings on ethics and consider a variety of problems in this area from a neuroanthropological perspective, unsettling long-held assumptions, providing novel ways of approaching old questions, and identifying new problems. These contributions show the potential depth and breadth of a neuroanthropology of ethics, as well as its distinctiveness. Future neuroanthropological work into ethics may also include, for example, studies investigating the implications of plural rationalities for brain function and ethical reasoning. Significant neuroethical questions in this context would include whether different forms of rationality [e.g., the elementary forms embedded in Douglas's (1978) and Thompson and colleagues' (1990) theory of socio-cultural variation or Fiske's (1991) relational models theory] are rooted in segregated neural mechanisms; under which circumstances are these mechanisms recruited; how do they constrain or bias moral reasoning; and to what extent they compete in guiding ethical reasoning and moral conduct.

An important area of development in the neuroanthropology of ethics involves the embedding of neuroscientific research in ethnographic fieldwork (Domínguez et al. 2010; Domínguez 2012; Immordino-Yang 2013). Neuroethnography, as this amalgamation of ethnography and neuroscience has been called (Domínguez et al. 2010), aims to provide an in-depth understanding of domains of cultural activity to better derive constructs, variables, and hypotheses that can be tested in the neuroscience laboratory; increase the ecological validity of neuroscience research; and shed light onto people's subjective experience including emotions, sense of self, sense of justice, preferences, and reasoning, which may be correlated with variations in brain activity and structure (cf. Immordino-Yang 2013). Neuroethnography can be a powerful tool in addressing neuroethical problems within the scope of the new anthropology of ethics, recentered as it is in the relational individual. These problems include those quoted at the outset such as how are ethical subjectivities formed, or how are they expressed in ordinary social encounters.

One domain of enquiry arousing an increasing amount of interest to which a neuroanthropology of ethics will be of relevance relates to research in neurorobotics, brain simulation, and neuromorphic computing aiming at artificially reproducing mental and behavioral properties of animals and humans. Research specifically aimed at developing robots with human (or human-like) attributes, simulating the human brain, and building autonomous devices with human-like learning and behavioral capabilities (cf. Glaskin 2012) will require a neuroanthropological understanding of what is to be human. The reason for this is that the human brain is a cultural brain and artificially reproducing mental and behavioral capabilities, which can be called truly human, will require an understanding of the human brain's mechanisms for acquiring culture and how these mechanisms interact with specific cultural milieus. Key among these capabilities will be those that allow robots, simulations, and devices to operate consistently with ethical codes of conduct.

A final fertile area of future investigation for the neuroanthropology of ethics concerns distributed cognition (see also ▶ Chap. 26, "Extended Mind and Identity"). Originally formulated by cognitive anthropologists, distributed cognition holds that "a society, as a group, might have some cognitive properties differing from those of the individual members of the group" (Hutchins 1991). From this perspective, cooperative activities, in which we may include ethical reasoning and decision making, require that a group of people form (to some extent) shared, coherent, and equivalent interpretations (Hutchins 1991). As argued earlier, the neural bases of moral agency are to be found beyond the confines of a single brain. Distributed cognition therefore offers a way of conceptualizing and investigating this central issue for a neuroanthropology of ethics. Levy (2007) has similarly remarked on the importance of distributed cognition for neuroethics as moral knowledge, he notes, is the product of the interaction of many individuals contributing different kinds and levels of expertise. Neuroanthropology can make important contributions to our understanding of the distributed character of moral knowledge, its production, and reproduction by, for example, hyperscanning during joint ethical judgment or reasoning tasks factoring in relevant cultural scripts and conventions; or by exploring the distribution of different types of neural representations pertaining to moral attitudes or conduct across a social group.

Conclusion

The papers in this section highlight the importance of neuroanthropology not only for a neuroscience of ethics (which has as object of study the neural basis of moral agency) but also for an ethics of neuroscience (which deals with the practice of neuroscience and its uses). Across both dimensions, these papers demonstrate the need for neuroethics to be sensitive to context, culture, history, and diversity as well as to the relationship between universalism and particularism, scientific fact and personal experience. Together with the areas of future research outlined above, these papers also offer new ways of approaching ethical phenomena and open up exciting new avenues of enquiry. Neuroanthropology is thus set to make a decisive contribution to our understanding of that which is the guiding light of our conduct.

Cross-References

▶ Consciousness and Agency
▶ Determinism and Its Relevance to the Free-Will Question
▶ Extended Mind and Identity
▶ Free Will and Experimental Philosophy: An Intervention
▶ Historical and Ethical Perspectives of Modern Neuroimaging
▶ Neuroethics of Neurodiversity
▶ Neuroimaging and Criminal Law
▶ Neuroimaging Neuroethics: Introduction
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play
▶ Neuroscience, Neuroethics, and the Media
▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering
▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Brun, C. C., Lepore, N., Pennec, X., Lee, A. D., Barysheva, M., Madsen, S. K., et al. (2009). Mapping the regional influence of genetics on brain structure variability: A tensor-based morphometry study. *NeuroImage, 48*, 37–49.

Chiao, J. Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., et al. (2010). Dynamic cultural influences on neural representations of the self. *Journal of Cognitive Neuroscience, 22*, 1–11.

Chiao, J. Y., Iidaka, T., Gordon, H. L., Nogawa, J., Bar, M., Aminoff, E., et al. (2008). Cultural specificity in amygdala response to fear faces. *Journal of Cognitive Neuroscience, 20*, 2167–2174.

D'Aquili, E. G., Laughlin, C. D., & McManus, J. (1979). *The spectrum of ritual: A biogenetic structural analysis*. New York: Columbia University Press.

Domínguez D, J. F. (2012). Neuroanthropology and the dialectical imperative. *Anthropological Theory, 12*, 5–27.
Domínguez D, J. F., Lewis, E. D., Turner, R., & Egan, G. F. (2009). The brain in culture and culture in the brain: A review of core issues in neuroanthropology. *Progress in Brain Research, 178*, 43–64.
Domínguez D, J. F., Turner, R., Lewis, E. D., & Egan, G. (2010). Neuroanthropology: A humanistic science for the study of the culture-brain nexus. *Social Cognitive and Affective Neuroscience, 5*, 138–147.
Douglas, M. (1978). *Cultural bias* (Occasional paper 35). London: Royal Anthropological Society.
Dumit, J. (2004). *Picturing personhood: Brain scans and biomedical identity*. Princeton: Princeton University Press.
Faubion, J. D. (2011). *An anthropology of ethics*. Cambridge: Cambridge University Press.
Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations*. New York: Free Press.
Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*, 814–821.
Glaskin, K. (2012). Empathy and the robot: A neuroanthropological analysis. *Annals of Anthropological Practice, 36*, 68–87.
Han, S., Mao, L., Gu, X., Zhu, Y., Ge, J., & Ma, Y. (2008). Neural consequences of religious belief on self-referential processing. *Social Neuroscience, 3*, 1–15.
Hedden, T., Ketay, S., Aron, A., Markus, H. R., & Gabrieli, J. D. (2008). Cultural influences on neural substrates of attentional control. *Psychological Science, 19*, 12–17.
Hutchins, E. (1991). The social organization of distributed cognition. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 283–307). Washington, DC: American Psychological Association.
Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., et al. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage, 21*, 1167–1173.
Immordino-Yang, M. H. (2013). Studying the effects of culture by integrating neuroscientific with ethnographic approaches. *Psychological Inquiry, 24*, 42–46.
Ketay, S., Aron, A., & Hedden, T. (2009). Culture and attention: Evidence from brain and behavior. *Progress in Brain Research, 178*, 79–92.
Kobayashi, C., Glover, G. H., & Temple, E. (2007). Children's and adults' neural bases of verbal and nonverbal 'theory of mind'. *Neuropsychologia, 45*, 1522–1532.
Kobayashi, C., Glover, G. H., & Temple, E. (2008). Switching language switches mind: Linguistic effects on developmental neural bases of 'Theory of Mind'. *Social Cognitive and Affective Neuroscience, 3*, 62–70.
Konvalinka, I., & Roepstorff, A. (2012). The two-brain approach: How can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience, 6*, 215.
Laidlaw, J. (2002). For an anthropology of ethics and freedom. *Journal of the Royal Anthropological Institute, 8*, 311–332.
Lambek, M. (2010). *Ordinary ethics: Anthropology, language, and action*. New York: Fordham University Press.
Lende, D. H., & Downey, G. (2012). Neuroanthropology and its applications: An introduction. *Annals of Anthropological Practice, 36*, 1–25.
Levy, N. (2007). *Neuroethics*. Cambridge: Cambridge University Press.
Lewis, R. S., Goto, S. G., & Kong, L. L. (2008). Culture and context: East Asian American and European American differences in P3 event-related potentials and self-construal. *Personality and Social Psychology Bulletin, 34*, 623–634.
Ma, Y., Bang, D., Wang, C., Allen, M., Frith, C., Roepstorff, A., et al. (2012). Sociocultural patterning of neural activity during self-reflection. *Social Cognitive and Affective Neuroscience*, *9*(1), 73–80.

Rilling, J. K., Barks, S. K., Parr, L. A., Preuss, T. M., Faber, T. L., Pagnoni, G., et al. (2007). A comparison of resting-state brain activity in humans and chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 17146–17151.
Roepstorff, A. (2002). Transforming subjects into objectivity. An ethnography of knowledge in a brain imaging laboratory. *Folk, Journal of the Danish Ethnographic Society, 44*, 145–170.
Thompson, M., Ellis, R., & Wildavsky, A. (1990). *Cultural theory*. Boulder: Westview Press.
Turner, R. (2012). The need for systematic ethnopsychology: The ontological status of mentalistic terminology. *Anthropological Theory, 12*, 29–42.
Whitehead, C. (2012). Why the behavioural sciences need the concept of the culture-ready brain. *Anthropological Theory, 12*, 43–71.
Whitehead, C., Marchant, J. L., Craik, D., & Frith, C. D. (2009). Neural correlates of observing pretend play in which one object is represented as another. *Social Cognitive and Affective Neuroscience, 4*, 369–378.
Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *NeuroImage, 34*, 1310–1316.
Zigon, J. (2008). *Morality: An anthropological perspective*. Oxford: Berg.

Justice: A Neuroanthropological Account

19

Charles D. Laughlin

Contents

Abstract

The central idea of this paper is that the directly experienced sense of justice and injustice is universal to peoples everywhere because the human brain is "wired" to experience empathy and harmony, and experience them associated with pleasure and injustice with pain. The sense of justice and injustice is shared with other big-brained social animals, particularly among social primates who share homologous neurophysiological structures with humanity. How the sense of justice plays out in societies depends upon neurocognitive, environmental,

C.D. Laughlin
Department of Sociology and Anthropology, Carleton University, Ottawa, ON, Canada
e-mail: cdlaughlin@gmail.com

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_135,

social structural, and cultural factors that impinge upon the development of each person's perceptions, values, and personality. What a people consider as real, how they conceive of causation, and the information they credence about events will certainly vary across cultures. The sense of justice, both for the individual and among individuals in a community, is a principle ingredient in social institutions among more acephalous societies, while it may vary considerably among more complex and hierarchical societies. It is likely that all societies recognize the lack of empathy and sense of justice/injustice among abnormal individuals, and that appropriate action must be taken to protect society from actions that unsettle the balance of justice. Finally, by applying A.M. Hocart's thinking to the issue, the individual's sense of justice and injustice may become alienated from social procedures and juridical institutions, especially in highly complex, demographically diverse, and bureaucratized states.

Introduction

In the early hours of March 13, 1964, bartender and manager Catherine Susan "Kitty" Genovese left Ev's 11th Hour Sports Bar in the Queens, New York City, at the end of her shift. Kitty, an attractive, petite 28-year-old Italian-American, got into her little red Fiat and drove the approximately five miles to her apartment in Kew Gardens which she shared with her partner, Mary Ann Zielonko. As it remains today, Kew Gardens was a quiet neighborhood, verdant with trees and lawns. When she arrived, she parked her car and stepped out into legend.

As Kitty walked toward her apartment, she was attacked by a man who chased her and stabbed her in the back. "Oh my God," she screamed, "he stabbed me! Help me! Please help me! Please help me!" (Ethier 2010, p. 25). Lights went on in apartments and the attacker ran away, only to return to the scene. By then, Kitty had crawled into a building where the attacker found her, raped, and robbed her and repeatedly stabbed her. The attack lasted around 30 min. Kitty died on the way to the hospital. During the attack, at least three people witnessed the stabbings and others heard her screams and saw her lying comatose on the ground. No one did anything to intervene, and only one person called the police. Because of a New York Times article that erroneously claimed that 38 people had witnessed the crime and had done nothing to help Kitty in her distress, the story went national and was followed by an outpouring of outrage. Since then, many of the "facts" of the case have been investigated and repudiated. What seems beyond dispute is that Kitty Genovese was killed by a serial homicidal psychopath, and that numerous people were reluctant to "get involved." The case has subsequently become the textbook example of what is now known as the "bystander effect," or the "Genovese syndrome" (Manning et al. 2007).

For present purposes, the exact facts of Kitty Genovese's murder are beside the point. What is germane here is the sense of moral outrage millions of people felt when they heard of the case. In the intervening half century between Kitty's tragic murder and the present, the reader may have experienced in some form this sense of

injustice, perhaps hundreds or thousands of times. Perhaps that sense has been experienced when someone is freed from prison having been found innocent due to new exculpatory evidence, as has happened scores of times since the development of DNA profiling. Perhaps it has been experienced when there is news of politicians and Catholic priests perverting the "course of justice" to their own ends. Or may be it is felt when a killer gets less than his/her due because of the common practice of plea bargaining. Moreover, just as many times the reader has experienced the opposite, a sense of justice having been done, as when Kitty's killer was caught, tried, and sentenced, or when a friend who has gotten the promotion he/she richly deserved after long years of work, or some dispute has been resolved equitably.

In this chapter, we will explore the phenomenology of the sense of justice and parse out the universal experiential elements of the sense. It will be shown that the core of the sense of justice is universal to peoples everywhere and underlies aspects of morality that pervade the lives of most normal people in every culture, regardless of their history (see Moghaddam 2010, p. 137). This sense of justice is universal, because it is how people's brains work. In other words, there are neurophysiological structures that all humans have in common and that are triggered under certain sets of circumstances. Because we are speaking of neurophysiological structures, we may presume that they have evolved, and that we may trace the anlage of the human sense of justice in other big-brained social mammals, especially among our close relatives, the chimpanzees. We will lodge what we know about the neuropsychology of justice in a psychodynamic theory of justice, and will examine how the sense of justice plays out among non-western traditional peoples around the globe. As we shall see, the closest most human languages and cultures come to our notions of "good" and "evil" gloss something like "harmony" and "disharmony," "wholeness," and "fragmentation," "constructive" and "destructive." The sense of justice is triggered in the presence of good and evil, and is produced by a complex entrainment of neurophysiological structures mediating pleasure linked to those mediating empathy, the aesthetic and archetypal sense of wholeness, and the judgment of fairness.

The Sense of Justice

The definitions of justice of the sort one encounter among western philosophers are of little use when asking either cross-cultural or evolutionary questions, for they are both ethnocentric and designed such that they only apply *in principle* to modern humans, as sociobiologists interested in the biology of morality have emphasized (Katz 2000b; Williams 1980, pp. 275–276). They are, in a word, scientific "dead ends." For instance, John Rawls (1971) argues that justice requires reason, and by reason, he means a certain range of human cognition. Moreover, he excludes emotion from a definition of justice, which, as we will see, is crucial to understanding justice as experienced by people everywhere and animals. That is why we emphasize the *sense of justice*. Robert Solomon, in his book

A Passion for Justice, notes that: "Justice presumes a personal concern for others. *It is first of all a sense*, not a rational or social construction, and I want to argue that this sense is, in an important sense, natural" (1995, p. 102; emphasis added). Elements of the experience of justice are universal (justice, universal aspects of), while the way different peoples *conceive* of justice and injustice will vary from local culture to local culture (McParland et al. 2011).

If we explore the etymology of the word, we see that justice has to do with things being perceived as right, exact, equitable, and fair. We are "justified" in our actions when we "get it right," we "follow the plan exactly," we "strike the right balance," we have "fair dealings" with one another. Simple enough, one would think, until one gets into protracted and torturous philosophical discourses on the nature and logic of justice, and then one faces a quagmire of hyper-rational theories that prove nearly useless for the purpose of grounding a neuroanthropology of morality, much less an neuroanthropological account of justice. It is perhaps far better that we ground our approach on the relatively simple sense of justice had by almost everyone, regardless of cultural background.

As Nythamar de Oliveira (2008, p. 124) notes, "For justice nowadays, more than ever, must be rethought in sustainable, phenomenological terms, as both its social, political and environmental, cosmological dimensions must correct the shortcomings of human, all-too-human *desiderata*." From this perspective, *injustice* seems to label the sense (perception, intuition, feeling, etc.) that something is not quite right, is distinctly wrong, off-kilter, unfair, inexact, unappealing, unbalanced, or injurious. (One thinks of the iconic image of Lady Justice carrying a balance scale and blindfolded against bias and special interests – sometimes a sword as well – an image that dates to the Roman goddess *Justitia*.) This sense is rarely if ever fully articulated or dispassionate. The sense of justice is inevitably accompanied by positive affect, injustice by negative affect. Thus, *the core of the sense of justice (justice, definition) is the empathetic pairing of positive feelings like pleasure, satisfaction, contentment, joy*, etc. *with the perception or intuition that someone's unfair/undeserved situation has been put to right, has been balanced, is now fair, and perhaps even healed* (same root as "whole" and "holy").

Justice and the Brain

The sense of justice is universal across cultures, and thus is an experience that must be mediated by inherent mechanisms in the brain (Laughlin 2011; Laughlin et al. 1990). This is yet another phenomenon amenable to a cultural phenomenological approach (see Csordas 1990, 1994; Throop 2009). However, for the purposes of explanation, we may use a relatively new strategy for accounting for such phenomena that has been developed in anthropology under the designation *cultural neurophenomenology* – a perspective that links the neurophysiology of direct experience with cultural factors (see Laughlin and Throop 2006, 2008, 2009). We humans are social primates, and our brains, like those of our primate relatives (see Russon 1996), are "wired" for social interactions and social learning by way of

mimicry. In other words, we are a "mimetic" animal – we are "wired" to learn by imitating our elders and peers (Garrels 2011). In their classic study, Meltzoff and Moore showed that newborns as early as 18-h old can reproduce tongue, mouth, facial, and other movements displayed by the adult they are observing (see Meltzoff and Moore 1977, 1983a, b, 1997; Meltzoff 1985, 2002). This proclivity for social mimicry is conditioned during our early development by both inherent neurogenesis and the style of nurture, actions, language(s), values, world view, and other cultural attributes of our parents and other caretakers. We are products of both our inherent neurophysiological mimetic nature and our culturally conditioned upbringing (Garrels 2011). The development of a person's sense of justice occurs in the context of this dialogue between neurocognitive mimetic growth and social conditioning. And as we have seen, the experience usually consists of a complex of elements, each one presumably mediated by discrete neurophysiological mechanisms.

Empathy

There is now a tremendous scientific interest in the capacity for and experience of empathy (or, sympathy; Wilson 1993), especially in anthropology (Hollan and Throop 2011; Throop 2008a, b, 2010a, b), primate ethology (De Waal 2009), forensic psychiatry (Baron-Cohen 2011), psychology (Haidt 2012; Farrow and Woodruff 2007), and neuroscience (Decety and Ickes 2009; Keysers 2011; Gallese 2003, 2005a, b). Of course, interest in the phenomenology of empathy dates back much further – at least to Edith Stein's 1917 classic *Zum Problem der Einfűhlung* (*On the Problem of Empathy*), a doctoral dissertation she completed under Edmund Husserl in 1916. This recent renewed interest arose on the tail of the discovery during the 1990s of mirror neurons by Giacomo Rizzolatti and his associates that appeared to lay the groundwork for a science of empathy (Braten 2007; Lepage and Théoret 2007; Iacoboni et al. 1999; Rizzolatti et al. 1996; Gellese et al. 1996).

The discovery of neural networks that are inherently "wired" to understand and mimic the actions of others cannot be overemphasized in the study of justice by either anthropologists studying morality or neuroethicists – especially those studies that have focused upon *mirror neurons* (Gallese 2005a; Gallese et al. 2007; Rochat et al. 2008, p. 231). It is apparently these systems that provide the structural foundation for empathy. By way of empathetic processing, a person may mimic the other's actions either directly or by imagining those actions and effects, and vicariously feel the other's joy, satisfaction, fear, pain, suffering, etc.

Empathy not only refers to action understanding and judgments pertaining to fairness; it also involves affect (Greene et al. 2001; Gallese et al. 2004). Wicker et al. (2003) have shown, using functional neuroimaging, that while mirror neurons mediate action understanding, they also mediate empathetic feelings. They tracked the activity of brain centers mediating the feeling of disgust in subjects that were stimulated with malodorous smells. Their facial gestures were

videoed and shown to other subjects in the absence of the smells. The same area (the so-called core disgust system: the anterior ventral insula which is located in the limbic system deep within the lateral sulcus) becomes active while viewing facial gestures signaling disgust as happens if one perceives a disgusting smell. Interestingly, direct electrical stimulation of the area will produce noxious sensations in the mouth. In addition, Amit and Greene (2012) have shown experimentally that if the sense of justice involves imagery, the affective influence upon moral judgments will likely be greater.

Fairness and Pleasure

As we have seen, the sense of justice is accompanied by positive affect and injustice by negative affect. In a series of experiments, Singer et al. (2006) have shown, using neuroimaging techniques, that perception of fairness is coupled with enhanced empathy and feelings of pleasure, while perception of unfairness is correlated with a decrease in empathy and pleasure. The areas involved in mediating the pleasure sense, or lack of it, include the nucleus accumbens (a structure that plays a central role in reward) and the orbito-frontal cortex (which mediates both decision-making and emotion/reward). The pleasure aspect of the sense of justice leads quite naturally to the interesting question of the aesthetics of justice (Ben-Dor 2011; Resnik and Curtis 2011). For instance, it is quite common for the sense of justice to be expressed in artistic forms cross-culturally (e.g., see Soroush 2007 on justice and beauty in Islam).

Empathy also involves feeling the suffering of others (Jackson et al. 2005; Botvinick et al. 2005). Focusing on the apperception of pain in others, Singer et al. (2004) have shown, using brain imaging, that a portion of the pain system is activated in the empathizing observer. The areas of the brain that light up when either feeling pain oneself, or observing pain in a loved one, are areas in the anterior insula, rostral anterior cingulate cortex, areas in the brain stem, and cerebellum.

Higher Centers of Empathy

Empathy, of course, involves more than affect and social identification with the pleasure and suffering of others (Katz 2000a). It also involves those areas of the brain that mediate social involvement, and especially those areas of cortex, like the prefrontal lobes, that modulate feelings and control attention and judgment. Seitz et al. (2006, p. 743), using surgical lesion data, have shown that: “These data complement the consistent observation that lesions of the medial prefrontal cortex interfere with a patient’s perception of own bodily state, emotional judgments, and spontaneous behavior. The results of the current meta-analysis suggest the medial prefrontal cortex mediates human empathy by virtue of a number of distinctive processing nodes.” Shamay-Tsoory et al. (2003), using a comparative lesion approach, have also isolated areas of the prefrontal

cortex – especially the right ventromedial prefrontal cortex – in the mediation of empathy. Among other things, the prefrontal cortex mediates our perception of the intentions of others, and is fundamental to social communication and interaction (Changeux 2002, Chap. 4).

Putting all of this together, the experience of empathy clearly has both an affective and cognitive aspect, both of which are critical to the sense of justice. Indeed, it is the presence of this deep-seated and inherent capacity for empathy that likely makes the difference between morality and mere cultural convention – a distinction that appears to also be universal (Blair and Blair 2011). Simon Baron-Cohen (2011, pp. 27–41) has similarly integrated many of these cortical and subcortical areas into what he calls the "empathy complex." His approach, however, includes more cognitive functions involving empathy than has been included in the present definition of the sense of justice, for his approach would by definition be limited to humans, rather than revealing the more universal and evolutionary elements of justice. What makes his approach especially interesting to us is that he shows how, by whatever psychological instrument used, the presence and intensity of empathy in any natural population of humans distributes along a bell curve, with most people falling in the mid-range and a few manifesting little or no empathy and a few manifesting intense affect (2011, pp. 20–27). It would be interesting to know if this kind of normal distribution occurs among other social animals. Moreover, it is becoming clear that the sense of justice, present in the child, undergoes a process of development in interaction between the child and his/her social environment (see Killen and de Waal 2000). The kinds of problem solving and judgments associated with the sense of justice range in complexity across individuals in any culture from the very concrete to the highly abstract (Rest et al. 1999).

"Wild" Justice

Among social animals, it makes sense to define justice in terms that fit with their situation: "Justice is a set of expectations about what one deserves and how one ought to be treated" (Bekoff and Pierce 2009, p. 113). In other words, animals have an inherent sense of fairness and normalcy, and react emotionally and even violently when these expectations are trampled on. Anyone with a pet dog will know how their beloved companion reacts when treated unfairly. So it is with human beings. Because the sense of justice is inherent in the structure of the human nervous system (see Wilson 1993), it must have evolved as part of a complex of neurophysiological features that mediate the psychological dimensions of sociality and sociability in our species, as well as other extinct species of hominins going back millions of years. Naturally, we have no direct phenomenological evidence of the sense of justice in any hominin species other than our own. *Homo habilis* and *Homo erectus* are long dead, and living chimpanzees and bonobos, like the family dog, have no language sufficiently complex to chat with us about their moral experiences. Yet it is reasonable to suppose that the anlage of our sense of justice

among earlier hominins were fairly similar to our own, perhaps without the level of self-reflection and cognitive complexity we associate with full-blown cultural theories of justice.

Inferring from their patterns of behavior and social interactions (dominance interactions, play, food-sharing, altruism, etc.), it is apparent that large brained social animals (justice, among social animals) do experience a sense of justice, as well as a spirit of cooperation and sharing (Bekoff 2007, Chap. 4; Bekoff and Pierce 2009; de Waal 2006, 2009; de Waal and Lanting 1997, p. 41; Lorenz 1964). Indeed, Marc Bekoff and Jessica Pierce make no bones about it. In their book, *Wild Justice* (2009, p. 1), they state: "Let's get right to the point... we argue that animals feel empathy for each other, treat one another fairly, cooperate towards common goals, and help each other out of trouble. We argue, in short, that animals have morality." With respect to justice, they go on to reason: "The principle of parsimony suggests the following hypothesis: a sense of justice is a continuous and evolved trait. And, as such, it has roots or correlates in closely related species or in species with similar patterns of social organization. It is likely, of course, that a sense of justice is going to be species-specific and may vary depending on the unique and defining social characteristics of a given group of animals; evolutionary continuity does not equate to sameness" (2009, p. 115).

We could present relevant data for a wide range of animal species, including elephants, dolphins, canids (including the domesticated dog) and various birds (see Bekoff and Pierce 2009; Peterson 2011; Hauser 2000 for broad discussions of morality among social animals), but our space is limited and it makes sense that we should stick to our primate relatives, especially the chimps and bonobos. Technically speaking, our closest living primate relative, determined by DNA data, is *Pan paniscus*, the bonobo (once called the "pygmy chimpanzee;" see de Waal and Lanting 1997; Lonsdorf et al. 2010; Wrangham et al. 1994). But *Pan troglodytes*, the "common" chimpanzee, is also close genetically and hence relevant. Also, much of the data suggest that the origins of human morality are to be found in processes widely distributed among social primates (Flack and de Waal 2000).

Sense of Justice in Chimps, Bonobos, and Other Primates

We are unable to directly "get into the mind" of another animal. Rather, we have to infer their experiences from available behavioral and neurophysiological data (see Russon et al. 1996; Cohen 2010; Aureli and de Waal 2000). Fieldwork among free-ranging primates and laboratory experiments suggest that empathy and the sense of justice is to be found during the activities and decision-making of primates involved in sharing (especially of food and grooming), the resolution of potential conflict, and in perception of and reaction to conditions leading to the experience of sympathy (or empathy).

The capacity for empathy can also manifest in the marked ability of chimpanzees to deceive others (Byrne and Whiten 1990). De Waal and Aureli (1996, p. 102)

offer an example of this phenomenon when Yeroen, a male chimpanzee living in the Arnhem Zoo Chimpanzee Colony in the Netherlands, was wounded by a rival male: “The wound was not deep and he walked normally when out of sight of his rival. However, for a week, Yeroen limped heavily each time he passed in front of his rival. If injuries inhibit further aggression, Yureon’s attempt to create a false image of pain and suffering may have served to delay renewed hostilities.” This kind of deception presumes that Yureon is capable of anticipating the feelings, perceptions, and judgments of the other. Chimpanzees and other primates, as well as other social mammals, stick up for their kin and friends during episodes of aggression and conflict, and chimps are famous for kissing and embracing each other after fights, a ritual enactment of empathy that would appear to calm ruffled feelings and reestablish peace (de Waal 2006, p. 19). Bonobos appear to be especially sensitive to others’ experiences, to what they may be feeling or intending, and what they need to do to keep relations peaceful and address their fellows’ suffering (de Waal and Lanting 1997, p. 154).

Justice in Cross-Cultural Perspective

It is perhaps redundant to emphasize that moral animals are social animals. It is doubtful that solitary species like the red fox experience a sense of justice, empathy, or any other moral affect/cognition neural entrainment for that matter. Hence, the manifestation of moral experience among humans usually plays out in a social context (McParland et al. 2011). Although the sense of justice, as we have seen above, is rooted in universal neural mechanisms, living in a family and community has an impact on how that sense manifests in different social environments. At the same time, there always exists a tension between the experience of individual persons and the social context, regardless of the society involved (O’Manique 2003, p. 133). The inevitable social context of the sense of justice and moral judgments generally is far more evident among non-technocratic, traditional societies than is the case in our own materialistic postindustrial society. This is why the study of morality and culture has occupied anthropologists for generations (see d’Andrade 1995; Throop 2009).

Justice in Polyphasic Cultures

One of the reasons for this disparity is how we in the west interpret experiences had in *alternative states of consciousness* (ASC) – experiences had while dreaming, pursuing vision quests, meditating, carrying out religious rituals, during drug trips, etc. In western technocratic societies, we are taught to ignore and even abjure experiences of this sort. We live in a *monophasic* (people ignore experiences had in other than “normal” waking consciousness) social environment (Laughlin 2011, pp. 62–66; Laughlin et al. 1990, pp. 154–156, 289–291; Laughlin and Throop 2001; Lumpkin 2001). The avoidance of socially meaningful interpretations of ASC is part and parcel of the conditioning necessary for the

extreme egoistic self-identity relished in the west, and incidentally for the pervasive sense of alienation in our society today.

Most non-western, non-materialistic human societies on the planet are *polyphasic* societies. Polyphasic cultures are markedly different from the ones to which most of us belong. These are societies in which experiences had in ASC are conceived by people as different domains of reality, not as unreality. Indeed, most people on the planet, even those in monophasic cultures, rarely if ever make a distinction between experienced reality and extramental reality in their everyday lives. Their sense of identity incorporates memories of experiences had in dreams and other ASC, as well as in waking life. Non-western people may in fact have no word in their native tongue that glosses precisely "dream" in our English sense. What we call a "dream" may be considered by others to be the polyphasic ego (soul, spirit, shadow, etc.) of a person experiencing another domain of reality during sleep. Dream experiences, just as waking experiences, inform the society's general system of knowledge about the self and the world, as well as the development of a person's identity. One can thus understand why ethnographer Jean-Guy Goulet's hosts among the Guajiro (a South American people) would not allow him to live with them unless he "knew how to dream" (1994, p. 22).

Why does the distinction between monophasic and polyphasic cultures matter vis-à-vis our discussion of justice (justice, in polyphasic cultures)? Because how a person interprets and acts upon the sense of justice (or injustice) depends in part upon how entangled one feels and knows himself/herself to be in the commonweal. Here in the mostly urban west, we can live in an apartment for years and not know many, if any of our neighbors. In traditional societies, everybody typically knows everybody else, has grown up with many of them, and moreover directly experiences the ineluctable interconnectedness between themselves and both the community and the physical environment. The experience of entanglement is heightened during dreams and other ASC where reality comes to be known as both physical (what one experiences in what we call waking consciousness) and spiritual (what one experiences in ASC when the normally invisible connections between things and events become manifest).

In traditional societies, such experiences lead to and confirm mythological accounts of the world. Mythological worldviews usually take the form of a cosmology in which everything is interconnected in space and time (Laughlin and Throop 2001). In such a cultural context, the view of every man "is an island" can hardly arise. Extreme egoism of the sort we encounter in modern technocratic societies is rare. Everybody's well-being is dependent upon everybody else's, as well as the well-being of spiritual beings and the cosmos in general. The relevance of this factor in traditional cultures will become clear in a moment.

Justice and Psychosocial Abnormalities

Considering that the sense of justice is universal to people with a normal brain, and that the empathetic aspect of that sense distributes along a normal curve

(see above), it is interesting to look at what this means to research into neuropsychological abnormalities in which little or no sense of justice or empathy is experienced. We know, for instance, that extreme autism is associated with what Vittorio Gallese (2006, p. 7) calls a lack of "intentional attunement" with others – an inability to internally construct a socially shared world full of others with whom one feels empathy. Gallese points at the likely disruption of mirror neuron networks during development among autistic individuals. Although otherwise quite dissimilar from autism, the same is likely true for psychopathy (Blair et al. 2005). Psychopathic individuals develop serious social and moral deficits, often due to early damage to their prefrontal cortex (Anderson et al. 1999; Saver and Damasio 1991), the area of the brain primarily responsible for socio-moral judgments and executive mediation of social affect. This leads to a profound empathetic dysfunction (Blair 2007) among individuals who may or may not perpetrate violent behaviors. Because of this dysfunction and associated low affect, psychopathic individuals very likely have little or no sense of justice – zoologically speaking, being more like the solitary red fox than the highly social arctic wolf. There is some evidence from state-level societies to suggest that psychopathy occurs across cultures (Cooke 1995, 1998; Shariat et al. 2010), but there have been distressingly few cross-cultural studies of psychopathy carried out among traditional peoples (Pitchford 2001).

However, it follows from the perspective developed here that one would expect to find virtually all societies recognize the lack of empathy or sense of justice in some of its members, and to conceive of, to speak about, and to act in respect to such individuals as abnormal (see also Moll et al. 2005). For instance, anthropologist Jane Murphy (1976) compared the Yoruba people of Nigeria with the Inuit people of Alaska with regard to how they conceive and respond to psychopathy-like behaviors and personality characteristics. Both societies recognize something akin to psychopathy, and consider the condition to be dangerous and abnormal. The Inuit label such individuals *kunlangeta*: "This is an abstract term for breaking of the many rules when awareness of the rules is not in question. It might be applied to a man who, for example, repeatedly lies and cheats and steals and does not go hunting and, when the other men are out of the village, takes sexual advantage of many women – someone who does not pay attention to reprimands and who is always being brought to the elders for punishment" (Murphy 1976, p. 1026). The typical reaction to such individuals was for the elders to kill him, for they considered the condition un-treatable.

Justice, Culture and the Law

It is thus not surprising that the sense of justice plays out in most traditional societies in social rituals – especially as ritual forms of adjudication that we westerners interpret as "courts" or "councils." (That is why most of the discussion of justice in anthropology centers around "law" and judicial institutions, and why the anthropology of law is a major ethnological subdiscipline

(see Miller 2001; Niezen 2010; Rosen 1989; Moore 2005; Donovan and Anderson 2003)). When the author was doing ethnographic fieldwork among the So – a pastoral and horticultural people living on three mountains in Northeastern Uganda – he transcribed a number of what we might call *moots* (Bohannan 1989, pp. 160–161) – a conflict brought before the council of elders (all male, the So being a patrilineal and patriarchal society). The social dimensions of the felt sense of justice and injustice represented in many of these moots are quite indicative of the kinds of juridical institutions for conflict resolution found among tribal peoples generally, and illustrate the extent to which the emotional factor in the sense of justice and injustice becomes clear. Moreover, the relationship between the sense of justice and social process is very likely similar to the prevailing situation going back tens of thousands of years, well back into the Upper Paleolithic.

To give a flavor of the kind of situation we anthropologists encounter in the field, the author will share this transcript taken from our ethnography of the So – a moot we will call the "Case of the Purloined Cow" (Laughlin and Allgeier 1979, pp. 128–131):

> A young man has come before the council accusing an old man (L and N, respectively) of having taken and killed one of his cows without permission. Just what N used the cow for is not clear. L made his accusation before one of the elders of the tribe who in turn called a meeting of all the elders the following morning. Approximately 75 elders were in attendance for the meeting. When all were gathered, L stood before the elders and made his accusation and also stated that N had repeatedly refused to repay the cow to him and that it had been over two years since the incident. L then sat down.
>
> Elder: Supports L's accusation (he and others apparently aware of the situation) and stated that in his opinion it was bad (*erono*) to kill the cow of another without the owner's permission.
>
> Elder: Addressing N, "Did you obtain L's permission before taking his cow?" (N replies in the negative.) "Then you must have your cow ready when the owner comes for repayment."
>
> N: "Why complain? I will payu the cow."
>
> Elder: "Keep quiet! I will speak. The man [L] is accusing you of killing his cow without permission. The question is whether you will pay him a cow in return. It is wrong (*erono*) that you did not ask."
>
> Elder: You [N] are guilty. When the owner of a cow you have taken asks for a cow in return, you should pay it."
>
> At this point an old woman (the mother of L sitting outside the circle of elders) interrupts to ask that the matter be taken directly to the police [i.e., the district office of the Uganda police]. She does not wish the elders to reach a decision here. She interrupts a number of times throughout the proceedings while remaining seated outside the council grounds (*omut*).
>
> Elder: "You two must talk kindly to each other and settle the giving of the cow. You must again ask N for a cow and if he again refuses you must accuse him again before us."
>
> L: "It is not bad to kill a cow of another if he knows it will be returned, but I have no assurance with this man." [L is visibly angry.]
>
> Elder: "Do not quarrel! You are killing this man from quarreling too much. Ask slowly for your cow."
>
> There is general outburst from various participants including L's mother. Some of the elders pound the ground with their walking sticks for order.

Elder: "You [N] should pay the cow today. It is not good for this matter to be taken to the government."

L's mother demands the matter be taken to the police.

Elder: "Do not say that! You will spoil the matter."

L's mother: "Take it to the police! We must have the cow quickly."

L: "I want my cow today!"

Elder: "You [N] must go today and select a cow for L. Do you [L] want a cow or a bull?" L indicates that he only wants a cow. "Will you [N] give this man a cow in the next while?"

N: "I am going to search somewhere for a cow."

A feeling is expressed on the part of some of the elders that the meeting should be postponed both because of the absence of the government-appointed chief (*Mukungu*) and because it will give N a chance to obtain a cow.

Elder: "Why are you saying postpone the meeting if the cow is not to be given today?"

Elder: "You [N] must tell in front of all the people here that you will pay the cow so all will know."

N remains silent.

Elder: "If we take this matter to the government, it will get bad, serious. We must settle this ourselves."

At this point several elders suggest that five days be given N to supply the cow. If it is not returned by that time, then all he elders will meet again.

L: "It is alright with me, but ask N."

N: "It is very difficult for me to find a cow unless I go to ask friends [meaning lineage relations] for one. Is it alright with you [L] if I go in search of this cow? I will consider this a serious matter."

L agrees.

Elder: "You can search for five days. Do not take this matter to the police. All of you [L and his party] go and wait for five days. If at the end of that time the cow is not replaced, we will gather again."

The replacement cow was indeed found among N's relatives and given back to L, thus bringing the matter to a close without appeal to external (and generally despised) government authority. Had the matter been taken to the police, the elders would have lost control, N would have been called a "thief" in English jurisprudence and likely jailed, and the collective sense that justice had been perceived to be done would have failed. From the So point of view, it was really a matter of balance or reciprocity of exchange between two groups of people, L's lineage (the lineage collectively, not the individual, own cattle) and N's lineage. There was another aspect that was enacted repeatedly in these proceedings – an intolerance of extreme negative emotion. When an elder intervenes in a quarrel and says "you are killing this man with your words, or accusations," *they mean it literally*. African cultures generally hold that to say something is tantamount to doing it. Thus, to get angry and threaten someone is taken very seriously. Every effort is made by the elders to "keep the lid on," to cool the emotions of disputants and avoid potential violence between individuals and feuds among lineages – eruptions that can easy lead to social disharmony.

As we mentioned above, the So are like most traditional peoples on the planet in that their sense of justice has a lot to do with perceived balance in matters of reciprocity. There is an understanding among such people that there is a natural balance with respect to both physical processes and social transactions. The sense of injustice among people is triggered when this balance is perceived to have been disrupted.

Reciprocity, Natural Balance, Healing, and Justice

The Navajo (Navajo Indians) are a group of horticulturalists (in some areas intensive agriculturalists) and sheep herders who inhabit the largest reservation in North America. Central to the Navajo world view and moral philosophy is the concept of *hozho* (or *hǿzhǿ*; see Witherspoon 1977; McNeley 1981; Farella 1984). The connotations of *hozho* are too broad to be precisely glossed in a single English word, but are usually translated as meaning "beauty," "peace," or "harmony." Perception of the loss of *hozho* in some circumstances is similar to, if not synonymous with, the sense of injustice (see Yazzie 1994, 1995). The emphasis in traditional Navajo culture is always toward reestablishing harmony (*hozho*), both within the individual and between individuals within the community. The Navajo term for holding a "justice ceremony" today is *Hozhooji Naat'aanii* (Bluehouse and Zion 1993) – borrowed from the ancient Chief Blessingway ceremony, a healing ceremony for reestablishing *hozho*.

As with all Native American societies, the Navajo have been impacted by Anglo values and institutions, and for a long time, their "criminal" matters were handled by western-style courts. Recent years have seen a return to traditional approaches to jurisprudence, and to the emphasis upon regaining harmony, peace, and healing (see Nielsen and Zion). Currently, most of the talk about "peace-making" and "restorative justice" (Sullivan and Tifft 2005; Nielsen and Zion 2005) emphasizes equality, reciprocity, and restoration of social harmony. In doing so, they conceive of the traditional approach to seeking justice to be diametrically opposite to that of the dominant Anglo "adversarial" systems. From the Navajo point of view, Anglo court proceedings often do not result in addressing the lived experience of injustice felt not only by the litigants, but also their families and the community at large. What is lost in Anglo proceedings is the recognition of the rippling waves of impact upon others. This recognition has more recently been acknowledged in Anglo jurisprudence with the introduction of victim impact statements at the end of the trial. But Navajo culture and peace courts fully acknowledge the interconnectedness of people with families and community as a whole, and the process keeps in mind as much as possible the sense of injustice felt by many who have been affected. In a word, the intent of the peace court process is the restoration of the sense of justice throughout the community.

In many ways, the way restorative justice plays out among the Navajo is representative of other Native American and Canadian First Nations peoples. The reification ("rediscovery") of traditional culture seems to happen with a good deal of editing (Povinelli 1993, p. 126). Along with emphasis upon restorative justice come problems in implementing systems forged in bygone, pre-conquest days into a more modern and socially complex situation. The Coast Salish people of British Columbia are a case in point, for in their distant past both peaceful and violent responses to disharmony routinely occurred, and not everyone was treated equally (Miller 2001, pp. 5–6). Moreover, the modern emphasis upon equality of social status belies the complex stratified political society characteristic of pre-Columbian days (Miller 2001, pp. 5–6).

A.M. Hocart and the Evolution of Justice

Political and cultural editing issues aside, it is beyond argument that individual societies range in the extent to which their institutions address peoples' sense of justice and injustice. Indeed, societies range on a continuum from those whose institutions prioritize their peoples' sense of justice to societies that do so minimally, if at all. In the latter case, this does not mean that people no longer manifest a sense of justice, but rather it is to whatever extent it is ignored in judicial proceedings. The demographically larger and more hierarchical and complex the justice system, the less likely are the institutions going to prioritize in favor of addressing that sense. Meanwhile, this sense, being universal to all peoples, remains present and is an ingredient in the efficacy of all local community systems. Some emphasis upon the therapeutic process seems to characterize most such systems.

In order to model the relationship between the inherent sense of justice/injustice and justice-oriented social institutions, it is useful to appeal to Hocart's theory of the evolution (justice, evolution of) of society. Arthur Maurice Hocart (Hocart, Arthur Maurice) (1883–1939) was a British sociocultural anthropologist living and working during the early decades of the twentieth century (1970[1936]). Hocart argued that all humans on the planet have at least one thing in common – *the quest for life*.

> Keeping alive is man's greatest preoccupation, the ultimate spring of all his actions. To keep alive he must have food. But man is not content merely to seize it and devour it. He looks ahead to ensure future supplies. Besides storage he also devised a technique for making food abound, a technique consisting of charms, spells, magic, and so on. (Hocart 1933, p. 133)

At some point in the ancient past, our ancestors became sufficiently self-reflexive to not only live, but *to conceive of life*. The quest for life among humans is a broader and more culturally nuanced occupation than merely assuring enough food and the procreation of the species – it involves the maintenance of vitality, longevity, equitable resource distribution, alliance among groups, comprehension and adaptation to the fact of death, and peaceful and harmonious social relations. The quest is essentially a social occupation in the sense both of cultural conditioning about what constitutes the good life, and the collective actions taken to assure the good life.

Hocart, being an anthropologist, realized that many peoples on the planet are not governed (Hocart 1970[1936], p. 128). And yet ungoverned peoples also seek the good life, value peace and harmony, and experience a sense of justice and injustice that must be addressed in a social context. The quest for life and for justice predates governance by untold thousands of years. How does this play out? For Hocart, the inherent need for life, for justice, leads to social solutions that, if they work satisfactorily, become institutionalized – hence the widespread institution of the moot we have described it above.

The Good, the Bad, and the Ugly: Justice and the Bureaucratic State

The genius of Hocart is obvious from his straightforward and simple insistence that the role of consciousness in human affairs is the biological, social, and cultural quest for the good life, and part and parcel of the good life is justice (Hocart 1934, 1970[1936], Chap. 10). People living in simple acephalous societies ("acephalous" means without a leader, but in Hocart's sense, the term implies an egalitarian society without a hierarchical system of governance) are no different than people living in complex hierarchical societies in that they all want the same fundamental thing – life. The furtherance of that universal desire may lead different peoples in different directions depending upon how they construct a technical and social adaptation to local contingencies. But underlying it all is a common, inexorable process of development. In the pursuit of the good life, the organization of the acephalous band and tribal society gradually develops a hierarchy – it loses that simpler horizontal organization typical of kin-based band, clan, moiety, and segmentary lineage systems and takes on the form of hierarchical officialdom. "Gradually the high rise higher, the low sink lower, until the state is rearranged in a vertical hierarchy such as ours" (Hocart 1970[1936]), p. 292). Within this pressure for hierarchy and coordination arises the process of *specialization of role*. People are taken into the government and taught to do jobs that no longer are carried out within and for the old kin-based system. Instead of disputes being settled by the elders with whom one lives, as in the So case above, the society develops specialized roles for jurists – individuals who develop a specialized knowledge of the law and procedure, but whom parties to the dispute may well not know.

Moreover, specialization becomes more and more constraining upon the individual's own quest for life. People in our modern technocratic society are forced to adapt to roles that no longer fulfill that primal desire – the quest for the good life. Much of the pressure is unavoidable, for it is driven by demographics – as the population grows, the competition for the means of attaining the good life, including justice, becomes more and more exclusionary and stressful. Somewhere along the line, the "coordinating system" (the bureaucratic state with its "criminal justice system," its system of codified laws and regulations, its amoral corporations, etc.) absorbs more and more of the resources available to the society – it becomes as it were a society within the society – and increasingly feeds its own needs, rather than the needs that gave rise historically to the institution in the first place. The coordinating system ceases to address the needs that initially gave rise to the system, but because it is now institutionalized – and worse, *bureaucratized* – it thwarts the existence of alternative institutions that might answer that primal need more directly and effectively. In the end, society becomes so fragmented and specialized that the result is the increasing alienation of the people who find their natural, innate quest for life – and justice (justice, alienation from the sense of) – thwarted and frustrated.

This alienation is nowhere more evident than in addressing the sense of justice/injustice, and the pursuit of social justice. The "peace court" restorative justice

movement among many aboriginal peoples is a tacit recognition of the experienced disjunction between the direct, individual and group quest for justice, and bureaucratically administered social processes.

Conclusion

Summarizing, the directly experienced sense of justice and injustice is universal to peoples everywhere. It is universal because the normal human brain is “wired” to experience empathy and harmony, and experience them associated with pleasure (and injustice with pain). We have seen that the sense of justice and injustice is shared with other big-brained social animals, particularly among social primates who share homologous neurophysiological structures with humanity. Just how the sense of justice plays out in societies depends upon neurocognitive, environmental, social structural, and cultural factors that impinge upon the development of each person’s perceptions, values, and personality. What a people consider to be real, how they conceive of causation, and the information they credence about events will certainly vary across cultures, especially when monophasic and polyphasic societies are compared. The sense of justice, both for the individual and among individuals in community, is a principal ingredient in social institutions among more acephalous societies, while it may vary considerably among more complex and hierarchical societies. It has been suggested that it is very likely that all societies recognize the lack of empathy and sense of justice/injustice among individuals with neuropsychological disorders, and that appropriate action must be taken to protect society from potential harm. Finally, by applying A.M. Hocart’s thinking to the issue, how the individual’s sense of justice and injustice may become alienated from social procedures and juridical institutions has been demonstrated, especially in highly complex, demographically diverse, and bureaucratized states.

People still experience a sense of justice and injustice while living in complex bureaucratic states, but whether that sense remains important, or is addressed directly by social justice institutions varies enormously across nation states and global, transnational juridical institutions (e.g., the Hague). Recognition of the failure to address the individual sense of justice has in recent decades led to the introduction of “victim impact statements” during the sentencing phase of major crimes in many countries and jurisdictions (see e.g., Erez and Tontodonato 1992; Bandes 1996). While this development has undoubted therapeutic effects for victims and their sense of justice, it is unclear whether such statements have any actual effect upon the judicial process (see Davis and Smith 1994; Erez et al. 2011).

When Kitty Genovese met her end on that early morning in 1964, it set the stage for a reflection upon the sense of justice and injustice felt by millions of people who had never met Kitty, as well upon the failure of social responsibility and the role of the modern justice system in addressing the sense of injustice felt by those who knew and cared for the victim. Nearly a half century has passed since these events,

and yet the issues of justice continue to reverberate through society. Kitty's killer is a psychopath, self-confessed serial rapist and murderer who remains in prison, despite numerous applications for parole. He is again qualified for a parole hearing in November 2013, and as always, Kitty's brother Vincent and other members of Kitty's family, as well as others impacted by the terrible crime, will be in the parole board room to see that he once again is denied parole and remains incarcerated. This speaks volumes about the reality and depth of the sense of justice and injustice, and how difficult it is for modern technocratic justice systems to directly address this inherent sense.

Cross-References

- ▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
- ▶ Consciousness and Agency
- ▶ Moral Cognition: Introduction
- ▶ Neurotheology
- ▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering
- ▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 23*(8), 861–868.

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience, 2*(11), 1032–1037.

Aureli, F., & de Waal, F. B. M. (Eds.). (2000). *Natural conflict resolution*. Berkeley: University of California Press.

Bandes, S. (1996). Empathy, narrative, and victim impact statements. *The University of Chicago Law Review, 63*(2), 361–412.

Baron-Cohen, S. (2011). *The science of evil: On empathy and the origins of cruelty*. New York: Basic Books.

Bekoff, M. (2007). *The emotional lives of animals: A leading scientist explores animal joy, sorrow, and empathy – And why they matter*. Novato: New World Library.

Bekoff, M., & Pierce, J. (2009). *Wild justice: The moral lives of animals*. Chicago: University of Chicago Press.

Ben-Dor, O. (2011). *Law and art: Justice, ethics and aesthetics*. Oxford: Routledge-Cavendish.

Blair, R. J. R. (2007). Empathic dysfunction in psychopathic individuals. In T. F. D. Farrow & P. W. R. Woodruff (Eds.), *Empathy in mental illness* (pp. 3–16). Cambridge: Cambridge University Press.

Blair, R. J. R., & Blair, K. S. (2011). Empathy, morality, and social convention: Evidence from the study of psychopathy and other psychiatric disorders. In J. Docety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 139–152). Cambridge, MA: MIT Press.

Blair, R. J. R., Mitchell, D., & Blair, K. (2005). *The psychopath: Emotion and the brain*. Oxford: Blackwell.

Bluehouse, P., & Zion, J. W. (1993). *Hozhooji Naut'aanii*: The Navajo justice and harmony ceremony. *Mediation Quarterly, 10*(4), 327–337. Reprinted in Nielsen, M. O., & Zion, J. W. (Eds.). (2005). *Navajo Nation peacemaking: Living traditional justice* (pp. 156–164). Tucson: University of Arizona Press.

Bohannan, P. (1989). *Justice and judgment among the Tiv*. Prospect Heights: Waveland.

Botvinick, M., Jha, A. P., Bylsma, L. M., Fabian, S. A., Solomon, P. E., & Prkachin, K. M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *NeuroImage, 25*, 315–319.

Braten, S. (2007). *On being moved: From mirror neurons to empathy*. Amsterdam: John Benjamins.

Byrne, R. W., & Whiten, A. (1990). Tactical deception in primates: The 1990 database. *Primate Report, 27*, 1–101.

Changeux, J.-P. (2002). The physiology of truth: Neuroscience and human knowledge. Cambridge, MA: Harvard University Press.

Cohen, J. (2010). *Almost chimpanzee: Searching for what makes us human, in rainforests, labs, sanctuaries, and zoos*. New York: Henry Holt.

Cooke, D. J. (1995). Psychopathic disturbance in the Scottish prison population: The cross-cultural generalizability of the hare psychopathy checklist. *Psychology, Crime & Law, 2*(2), 101–118.

Cooke, D. J. (1998). Psychopathy across cultures. In Psychopathy: Theory, research and implications for society (pp. 13–45). Springer, Netherlands.

Csordas, T. J. (1990). Embodiment as a paradigm for anthropology. *Ethos, 18*, 5–47.

Csordas, T. J. (Ed.). (1994). *Embodiment and experience: The existential ground of culture and self*. Cambridge: Cambridge University Press.

D'Andrade, R. (1995). Moral models in anthropology. *Current Anthropology, 36*(3), 399–406.

Davis, R. C., & Smith, B. E. (1994). The effects of victim impact statements on sentencing decisions: A test in an urban setting. *Justice Quarterly, 11*(3), 453–469.

De Oliveira, N. (2008). Husserl, Heidegger, and the task of a phenomenology of justice. *Veritas, 53*(1), 123–144.

De Waal, F. B. M. (2006). *Primates and philosophers: How morality evolved*. Princeton: Princeton University Press.

De Waal, F. B. M. (2009). *The age of empathy: Nature's lessons for a kinder society*. New York: Three Rivers.

De Waal, F. B. M., & Aureli, F. (1996). Consolation, reconciliation, and a possible cognitive difference between macaque and chimpanzee. In A. E. Russon, K. A. Bard, & S. T. Parker (Eds.). Reaching into thought: The minds of the great apes (pp. 80–110). Cambridge: Cambridge University Press.

De Waal, F., & Lanting, F. (1997). *Bonobo: The forgotten ape*. Berkeley: University of California Press.

Decety, J., & Ickes, W. (Eds.). (2009). *The social neuroscience of empathy*. Cambridge, MA: MIT Press.

Donovan, J. M., & Anderson, H. E. (2003). *Anthropology and law*. New York: Berghahn.

Erez, E., & Tontodonato, P. (1992). Victim participation in sentencing and satisfaction with justice. *Justice Quarterly, 9*(3), 393–417.

Erez, E., Kilchling, M., & Wemmers, J. (Eds.). (2011). *Therapeutic jurisprudence and victim participation in justice: International perspectives*. Durham: Carolina Academic Press.

Ethier, B. (2010). *True crimes: New York City*. Mechanicsburg: Stackpole.

Farella, J. R. (1984). *The main stalk: A synthesis of Navajo philosophy*. Tucson: University of Arizona Press.

Farrow, T., & Woodruff, P. W. R. (Eds.). (2007). *Empathy in mental illness*. Cambridge: Cambridge University Press.

Flack, J. C., & de Waal, F. B. M. (2000). 'Any animal whatever': Darwinian building blocks of morality in monkeys and apes. In L. D. Katz (Ed.), *Evolutionary origins of morality* (pp. 1–29). Bowling Green: Imprint Academic.

Gallese, V. (2003). The manifold nature of interpersonal relations: The quest for a common mechanism. *Philosophical Transactions of the Royal Society of London B, 358*, 517–528.

Gallese, V. (2005a). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences, 4*, 23–48.

Gallese, V. (2005b). "Being like me": Self–other identity, mirror neurons and empathy. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From cognitive neuroscience to social science* (Vol. 1, pp. 101–118). Cambridge, MA: MIT Press.

Gallese, V. (2006). Intentional attunement: A neurophysiological perspective on social cognition and its disruption in autism. *Brain Research, 1079*, 15–24.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Science, 8*, 396–403.

Gallese, V., Eagle, M. N., & Migone, P. (2007). Intentional attunement: Mirror neurons and the neural underpinnings of interpersonal relations. *Journal of the American Psychoanalytic Association, 55*, 131–176.

Garrels, S. R. (Ed.). (2011). *Mimesis and science: Empirical research on imitation and the mimetic theory of culture and religion*. East Lansing: Michigan State University Press.

Goulet, J.-G. (1994). Dreams and visions in other lifeworlds. In D. E. Young & J.-G. Goulet (Eds.), *Being changed by cross-cultural encounters* (pp. 16–38). Peterborough: Broadview Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Pantheon.

Hauser, M. D. (2000). *Wild minds: What animals really think*. New York: Henry Holt.

Hocart, A. M. (1933). *The progress of man: A short survey of his evolution, his customs and his works*. London: Methuen.

Hocart, A. M. (1934). The role of consciousness in evolution. *Psyche Annual, 14*, 160–164.

Hocart, A. M. (1970[1936]). *Kings and councillors: An essay in the comparative anatomy of human society*. Chicago: University of Chicago Press.

Hollan, D. W., & Throop, C. J. (Eds.). (2011). *The anthropology of empathy: Experiencing the lives of others in pacific societies*. New York: Berghahn.

Iacoboni, M., et al. (1999). Cortical mechanisms of human imitation. *Science, 286*(5449), 2526–2528.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others: A window into the neural processes involved in empathy. *NeuroImage, 24*, 771–779.

Katz, L. D. (2000a). Toward good and evil: Evolutionary approaches to aspects of human morality. In L. D. Katz (Ed.), *Evolutionary origins of morality* (pp. ix–xvi). Bowling Green: Imprint Academic.

Katz, L. D. (Ed.). (2000b). *Evolutionary origins of morality*. Bowling Green: Imprint Academic.

Keysers, C. (2011). *The empathic brain: How the discovery of mirror neurons changes our understanding of human nature*. Amsterdam: Christian Keysers.

Killen, M., & de Waal, F. B. M. (2000). The evolution and development of morality. In M. Killen & F. B. M. de Waal (Eds.), *Natural conflict resolution* (pp. 352–374). Berkeley: University of California Press.

Laughlin, C. D. (2011). *Communing with the gods: Dream cultures and the dreaming brain*. Brisbane: Daily Grail.

Laughlin, C. D., & Allgeier, E. R. (1979). *An ethnography of the So of northeastern Uganda*. New Haven: Human Relations Area Files Press.

Laughlin, C. D., & Throop, C. J. (2001). Imagination and reality: On the relations between myth, consciousness, and the quantum Sea. *Zygon, 36*(4), 709–736.

Laughlin, C. D., & Throop, C. J. (2006). Cultural neurophenomenology: Integrating experience, culture and reality through fisher information. *Journal Culture & Psychology, 12*(3), 305–337.
Laughlin, C. D., & Throop, C. J. (2008). Continuity, causation and cyclicity: A cultural neurophenomenology of time-consciousness. *Time & Mind, 1*(2), 159–186.
Laughlin, C. D., & Throop, C. J. (2009). Husserlian meditations and anthropological reflections: Toward a cultural neurophenomenology of experience and reality. *Anthropology of Consciousness, 20*(2), 130–170.
Laughlin, C. D., McManus, J., & d'Aquili, E. G. (1990). *Brain, symbol and experience: Toward a neurophenomenology of human consciousness*. New York: Columbia University Press.
Lepage, J. F., & Théoret, H. (2007). The mirror neuron system: Grasping other's actions from birth? *Developmental Science, 10*(5), 513–529.
Lonsdorf, E. V., Ross, S. R., & Matsuzawa, T. (Eds.). (2010). *The mind of the chimpanzee: Ecological and experimental perspectives*. Chicago: University of Chicago Press.
Lorenz, K. (1964). Moral-analoges Verhalten der Tiere-Erkenntnisse heutiger Verhaltensforschung. *Universitas, 19*, 43–54.
Lumpkin, T. W. (2001). Perceptual diversity: Is polyphasic consciousness necessary for global survival? *Anthropology of Consciousness, 12*(1–2), 37–70.
Manning, R., Levine, M., & Collins, A. (2007). The Kitty Genovese murder and the social psychology of helping: The parable of the 38 witnesses. *American Psychologist, 62*(6), 555–562.
McNeley, J. K. (1981). *Holy wind in Navajo philosophy*. Tucson: University of Arizona Press.
McParland, J., Eccleston, C., Osborn, M., & Hezseltine, L. (2011). It's not fair: An interpretative phenomenological analysis of discourses of justice and fairness in chronic pain. *Health, 15*(5), 459–474.
Meltzoff, A. N. (1985). The roots of social and cognitive development: Models of Man's original nature. In T. M. Field & N. A. Fox (Eds.), *Social perception in infants*. Norwood: Ablex Publishing.
Meltzoff, A. N. (2002). Elements of a developmental theory of imitation. In W. Prinz & A. Meltzoff (Eds.), *The imitative mind: Development, evolution and brain bases* (pp. 19–41). Cambridge: Cambridge University Press.
Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78.
Meltzoff, A. N., & Moore, M. K. (1983a). Newborn infants imitate adult facial gestures. *Child Development, 54*, 702–709.
Meltzoff, A. N., & Moore, M. K. (1983b). The origins of imitation in infancy: Paradigm, phenomena, and theories. In L. P. Lipsitt (Ed.), *Advances in infancy research* (Vol. 2). Norwood: Ablex.
Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6*, 179–192.
Miller, B. G. (2001). *The problem of justice: Tradition and law in the Coast Salish world*. Lincoln: University of Nebraska Press.
Moghaddam, F. M. (2010). *The new global insecurity*. Santa Barbara: Praeger.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*, 799–809.
Moore, S. F. (Ed.). (2005). *Law and anthropology: A reader*. Oxford: Blackwell.
Murphy, J. M. (1976). Psychiatric labeling in cross-cultural perspective: Similar kinds of disturbed behavior appear to be labeled abnormal in diverse cultures. *Science, 191*(4231), 1019–1028.
Nielsen, M. O., & Zion, J. W. (Eds.). (2005). *Navajo Nation peacemaking: Living traditional justice*. Tucson: University of Arizona Press.
Niezen, R. (2010). *Public justice and the anthropology of law*. Cambridge: Cambridge University Press.

O'Manique, J. (2003). *The origins of justice: The evolution of morality, human rights, and the law*. Philadelphia: University of Pennsylvania Press.

Peterson, D. (2011). *The moral lives of animals*. New York: Bloomsbury.

Pitchford, I. (2001). The origins of violence: Is psychopathy an adaptation? *Human Nature Review, 1*, 28–36.

Povinelli, E. A. (1993). *Labor's lot: The power, history and culture of aboriginal action*. Chicago: University of Chicago Press.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Resnik, J., & Curtis, D. E. (2011). *Representing justice: The creation and fragility of courts in democracies*. New Haven: Yale University Press.

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: A Neo-Kohlbergian approach*. Mahwah: Lawrence Erlbaum.

Rizzolatti, G., et al. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141.

Rochat, M. J., Serra, E., Fadiga, L., & Gallese, V. (2008). The evolution of social cognition: Goal familiarity shapes monkeys' action understanding. *Current Biology, 18*, 227–232.

Rosen, L. (1989). *The anthropology of justice: Law as culture in Islamic society*. Cambridge: Cambridge University Press.

Russon, A. E. (1996). Imitation in everyday use: Matching and rehearsal in the spontaneous imitation of rehabilitant orangutans. In A. E. Russon, K. A. Bard, & S. T. Parker (Eds.), *Reaching into thought: The minds of the great apes* (pp. 152–176). Cambridge: Cambridge University Press.

Russon, A. E., Bard, K. A., & Parker, S. T. (Eds.). (1996). *Reaching into thought: The minds of the great apes*. Cambridge: Cambridge University Press.

Saver, J. L., & Damasio, A. R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia, 29*(12), 1241–1249.

Seitz, R. J., Nickel, J., & Azari, N. P. (2006). Functional modularity of the medial prefrontal cortex: Involvement in human empathy. *Neuropsychology, 20*(6), 743–751.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: The role of the right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience, 15*(3), 324–337.

Shariat, S. V., Assadi, S. M., Noroozian, M., Pakravannejad, M., Yahyazadeh, O., Aghayan, S., Michie, C., & Cooke, D. (2010). Psychopathy in Iran: A cross-cultural study. *Journal of Personality Disorders, 24*(5), 676–691.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162.

Singer, T., Seymour, B., O'Doherty, J., Stephan, K. L., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*, 466–469.

Solomon, R. C. (1995). *A passion for justice: Emotions and the origins of the social contract*. Lanham: Rowman and Littlefield.

Soroush, A. (2007). The beauty of justice. *Centre for the Study of Democracy Bulletin, 14*(1 & 2), 8–12.

Sullivan, D., & Tifft, L. (2005). *Restorative justice: Healing the foundations of our everyday lives* (2nd ed.). Monsey: Willow Tree Press.

Throop, C. J. (2008a). On the problem of empathy: The case of Yap (Waqab), Federated States of Micronesia. *Ethos, 36*(4), 402–426.

Throop, C. J. (2008b). From pain to virtue: Dysphoric sensations and moral sensibilities in Yap (Waqab), Federated States of Micronesia. *Transcultural Psychiatry, 45*, 253–286.

Throop, C. J. (2009). 'Becoming beautiful in the dance': On the formation of ethical modalities in Yap, Federated States of Micronesia. *Oceania, 79*, 179–201.

Throop, C. J. (2010a). *Suffering and sentiment: Exploring the vicissitudes of experience and pain in Yap*. Berkeley: University of California Press.

Throop, C. J. (2010b). Latitudes of loss: On the vicissitudes of empathy. *American Ethnologist, 37*(4), 281–282.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron, 40*, 655–664.

Williams, B. A. O. (1980). Conclusion. In G. S. Stent (Ed.), *Morality as a biological phenomenon: The presuppositions of sociobiological research* (pp. 275–285). Berkeley: University of California Press.

Wilson, J. Q. (1993). *The moral sense*. New York: Free Press.

Witherspoon, G. (1977). *Language and art in the Navajo universe*. Ann Arbor: University of Michigan Press.

Wrangham, R. W., de McGrew, W. C., Waal, F. B. M., & Heltne, P. G. (Eds.). (1994). *Chimpanzee cultures*. Cambridge, MA: Harvard University Press.

Yazzie, R. (1994). 'Life comes from it': Navajo justice concepts. *New Mexico Law Review, 24*, 175–190. Reprinted in Nielsen, M. O., & Zion, J. W. (Eds.). (2005). *Navajo Nation peacemaking: Living traditional justice* (pp. 42–58). Tucson: University of Arizona Press.

Yazzie, R. (1995). Healing as justice: The American experience. *Justice as Healing: A Newsletter on Aboriginal Concepts of Justice* (pp. 1–5). Spring issue.

Free Will, Agency, and the Cultural, Reflexive Brain

20

Stephen Reyna

Contents

Abstract

This chapter explores free will, agency, and the brain. It employs a critical structural realist (CSR) approach to assist investigators in the study of will and of the possibilities for freedom. The position is argued in three sections. First, there is a preliminary discussion of an influential modern concept of will that identifies one of its weaknesses. The section concludes with an introduction to CSR. The following section addresses this weakness by rethinking will in CSR terms as the operation of a cultural, reflexive human brain. The following section speculates upon the implications of such an understanding of will for freedom. At this juncture, a notion of agency is offered and its relevance to free will is discussed.

S. Reyna
Max Planck Institute of Social Anthropology, Halle, Germany

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_138,

Introduction

A worthy essay should end with a bang – a eureka moment – when readers expostulate, "wow, that's it!" This chapter considers free will, agency, and the brain. It utilizes a critical structural realist (CSR) approach (Reyna 2002, 2012) to aid observers to investigate the human will and its possibilities for freedom. This position is formulated in three sections. First, there is a preliminary discussion of an influential modern conception of will, identifying one of its key weaknesses. This section concludes with an introduction to CSR. The following section addresses this weakness by rethinking will in CSR terms as the operation of a cultural, reflexive human brain. The final section speculates upon the implications of such an understanding of will for freedom. At this juncture, a notion of agency is offered, which leads to something of a eureka moment in the conclusion.

Inquiry concerning human will and the possibility of its freedom has exercised Western thought since the Greeks. Most of this speculation has been philosophic in nature. CRS is a view that reality is knowable to different degrees; that it consists overwhelmingly of structures; and that knowledge of these is needed for the formulation of normative judgment. An advantage of CRS is that it focuses attention upon something neglected in philosophical accounts of free will, the material structure of will. But in order to develop a CSR approach to free will, let us first consider a central, modern view of this construct.

Something Missing?

A Modern Problematic of Will

The philosophic problematic of will [will, functional accounts of] emerged during the Enlightenment and greatly developed over the nineteenth century. Two key Enlightenment figures of this problematic were David Hume and Immanuel Kant. By will Hume (1739–1740) meant "nothing but the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body, or perceptions of our mind" (1739–1740, p. 399). Will for Hume is something internal to humans that "knowingly" gives "rise" to the body's motions, i.e., actions, and/or its mental perceptions, i.e., thought. Hume informs his readers what will does, and in so doing gives a functional account of it, but not what the something is performing will's functions. Nowhere does Hume explicate what will is.

Hume was the culmination of British eighteenth century empiricism. Kant, working several decades later on a project to synthesize empiricism and idealism, operated with a notion of will influenced by Hume. Kant's will [will, as practical reason] (1785 [1997]; 1788 [1956]) is practical reason. As with Hume, Kantian will is internal to humans. It operates by a reasoning that applies a general principle of action pertaining to ethical duties to one's particular situation, because "Who wills the end, wills (so far as reason has decisive influence on his action) also the means

which are indispensably necessary and in his power" (1785 [1997/1949], pp. 84–85). So for Kant, as with Hume, will "knowingly" – specifically through reason – gives rise to action. Further, Kant's understanding of will is functional like Hume's, explicating what will does. There is talk of "faculties" and "categories" in Kant's work, but the something that uses these to execute the function of practical reason is passed over in silence.

The Enlightenment was followed by a rise of Romanticism and a decentering of reason's hold over human action. Arthur Schopenhauer, who imagined himself a follower and critic of Kant (see especially Schopenhauer 2000 [1818]), was an early nineteenth century exponent of this trend. About will he affirmed,

> In my language, this means that the objective world, the world as representation, is not the only side of the world, but merely its external side, so to speak, and that the world has an entirely different side, which is its innermost being, its kernel, the thing-in-itself. This we shall consider . . . 'will'. . . (2000 [1818/1969], p. 31).

Schopenhauer's will resembles that of Hume and Kant, being something internal to humans, their "innermost being." But what is this being? Troxell, a student of the philosopher, puts it as follows:

> [Schopenhauer's] . . . will is a blind, unconscious force that is present in all of nature. Only in its highest objectifications, that is, only in animals, does this blind force become conscious of its own activity. Although the conscious purposive striving that the term 'will' implies is not a fundamental feature of the will, conscious purposive striving is the manner in which we experience it and Schopenhauer chooses the term with this fact in mind (2011, p. 2).

For Schopenhauer, this "blind force" manifests itself in humans as impulse, instinct, and craving towards an insatiable goal, a "will to live" [will, as will to live], a purpose. Clearly, will governs in Schopenhauer's ontology – "a blind unconscious force" universally present in reality. Humans' experience this force as "conscious purposive striving." Again, as with both Hume and Kant, will is conceptualized functionally in what it does, "purposive striving." However, one ransacks Schopenhauer's work in vain looking for the "thing-in-itself" that does the striving.

Schopenhauer inclined a number of thinkers to his views at the end of the nineteenth century. Important among those Schopenhauer influenced was Sigmund Freud, who largely based his philosophic psychology on the pleasure principle: *It seems that our entire psychical activity is bent upon* procuring pleasure *and* avoiding pain*, that it is automatically regulated by the Pleasure Principle* (1935, p. 311, emphasis in the original). "Psychical activity" is "*essentially unconscious*" involving "*instinctive forces*" (1935, pp. 22, 24). Freud follows Schopenhauer to the extent that his will is a striving to achieve an end, with Freud's goal a will to pleasure [will, as will to pleasure] and Schopenhauer's a will to live. Freud explicitly recognizes that this understanding of will is functional, stating that the pleasure principle "subserves a certain function" (1950, p. 144). Freud deploys a lively pantheon of concepts – ego, id, superego, unconscious, transference, repression, etc. – to explain how humans went about expressing their will to pleasure, though he never identified the something that these concepts represented.

An equally significant follower of Schopenhauer, though equally a critic, was Friedrich Nietzsche. He was anti-Freudian in his insistence that, "man does not seek pleasure..." (1901 [1968], p. 273). He was anti-Schopenhauer as to the goal of his will asserting:

> Suppose, finally, we succeeded in explaining our entire instinctive life as the development and ramification of one basic form of the will–namely, of the will to power, as my proposition has it... then one would have gained the right to determine all efficient force univocally as–will to power. The world viewed from inside... it would be "will to power" and nothing else (1886 [1989], p. 36).

Humanity's goal for Nietzsche was not life or pleasure, but power. But he was pro-Schopenhauer in emphasizing the ontological importance of this end, insisting that "...the innermost being is will to power" [will, as will to power] (1901 [1968]); and crying out in the last two lines, "This world is the will to power–and nothing besides! *And you yourselves are also this will to power –and nothing besides!*" (1901 [1968], p. 550, emphasis in the original). As with the other philosophers, will is conceived functionally, in terms of what it does, which is to seek "power – and nothing besides!" Critically, one can search Nietzsche's work and find nothing of the something that makes the will to power.

Hume, Kant, Schopenhauer, Freud, and Nietzsche were not the only thinkers to conceptualize the notion of will, but they were arguably its iconic scholars; they created what might be called a modern problematic of will [will, modern problematic of] (MPW), composed of five similarities and one difference. The first similarity concerned location. Will was something within humans. It was "innermost being." Second, will was understood functionally, as something that innermost being does. The third and fourth shared characteristics pertained to the will's functions. The third commonality was that will produced action, and the fourth was that action was guided to purposes, ends, or goals. Our scholars debated these ends. They could be for pleasure, life, or power; but actions always strove in a particular direction. Before proceeding to the final similarity, consider the main difference. The Enlightenment thinkers imagined will formed action in some way out of knowing. Their post-Enlightenment counterparts had action fashioned out of something(s) other than knowing – be they impulse, craving, or instinct. The MPW conceptualizes that will as something that functions to produce purposive action in humans, through rational or non-rational means.

This brings us to a final commonality, which was an absence, with implications for free will. The MPW philosophers had nothing to say about the something that produced will. This is like saying the heart functions to circulate blood, without examining the organ that is the heart itself. The preceding has a major implication for free will. Scholars for millennia have pondered free will but, if the MPW is some sort of culmination of this tradition, it is clear that scholars cannot judge whether will is, or is not, free, because they literally do not know what it is that might, or might not, be free. Further complexities arise when one begins to examine the concept of freedom as it pertains to will.

Free Will and Critical Structural Realism

Consider the Oxford English Dictionary's definition of "Free" as "…actions, activity, motion, etc.: Unimpeded, unrestrained, unrestricted, unhampered" (1989, p. 158). A test of whether there is *unfree* will [will, restrictions on] is observation that a person's will is impeded, restrained, restricted, or hampered, i.e., whether there are "restrictors" upon will. Two categories of restrictors might be imagined – those that are internal and those that are external to persons. Certainly, it is my will to live forever. Deplorably, the aging process within my antique body will kill me off, if something else does not get me first. Certainly, I have always wanted to be a most brilliant mathematician. Lamentably, I had the lowest of low scores in math testing; something inside of me makes it impossible for me to grasp the simplest of mathematical concepts. Here, then, are two examples of internal restrictors of will. Certainly, it is my will to have a nice, three-bedroom apartment on Gramercy Park in New York City. Regrettably, my income as a professor inhibits realization of this will. Certainly, it is my will to have a more scientific anthropology. Deplorably, this is a time of postmodern anthropology, when my will goes frustrated. Here, then, are two external restrictors of will.

On the other hand, by my sophomore year in college, it became my will to become an anthropologist. Luckily, whatever was inside me intellectually allowed me to become an anthropologist, and the easily accessible funding and the availability of jobs allowed me to realize this will. The intellectual capacity to think anthropologically plus the existence of funding and job opportunities might be thought of as free will's internal and external "expediters." These observations of the vicissitudes of my will lead to the following conclusion: Sometimes the will is free to achieve its goals, sometimes it is not; i.e., there is and there is not free will, whatever will might be. The preceding discussion suggests that two sorts of inquiries are pertinent to a project analyzing will and its *soi disant* freedom. The first is the study of the something itself that is will. The second is discovery of the restrictors and expediters of that will, with the goal of eliminating restrictors and enhancing expediters. Below it is explained how critical structural realism can contribute to this project.

Critical structural realism (CSR) is an epistemology, ontology, and political ethics (Reyna 2002, 2012). Epistemologically, it argues for a realist understanding of knowledge in which science, fortified by elimination of Comtean positivist hubris, is a better tool for knowing about reality than any of the others. Ontologically, the knowledge it seeks is about structures of being and the processes (or logics) of their operation. Further, it is interested in understanding connections between structures: for example, relations between nonhuman and human structures as well as relations between human social organizations and those within individuals. A prominent circumstance of humanity is that people are suspended in webs of structure. So, as Horkheimer put it, this knowledge of these webs of structure is needed "…to liberate human beings from the circumstances that

enslave them" (1982, p. 244). This chapter pertains more to the structural aspect of CSR, because it is concerned to determine the material being that is will.

CSR is interested in systems of structures, that is, with the organization of forms and the interrelationships of these organizations. Two general categories of structures are distinguished – those in I-space and those in E-space. The former are structures internal to the human body and the latter are those external to the body. "I-space" structures include those of the respiratory system, etc. The key I-space structure under examination is that of the nervous system, especially its central organ, the brain. There are two major categories of "E-space" structures – those of human organizations (social systems) and those of nonhuman organizations (natural systems). It is judged that in some way I- and E-space structures are connected and that a central chore of CSR is to discover the nature of these connections. The main claim I want to make is that due to its connections with E-space, and what it does with these connections, the brain, and its activities, is the I-space structure that is the missing something that is the human will. The next section provides a sketch of the something that is the will.

Something Found: The Reflexive Cultural Brain

Thinkers working within the tradition of MPW theorized will functionally rather than structurally as CSR would prescribe. This is not altogether surprising as the structure in I-space that is key to understanding human behavior and cognition from a contemporary perspective, the brain, has been exceptionally difficult to study in operation to find out how its parts work together to generate function. However, this is now being reversed, particularly with the introduction of neuroimaging techniques (e.g., positron emission tomography, magnetic resonance imagining, and magnetoencephalography) that allow observation of the brain as it goes about its business. Consequently, will and the possibility of free will have begun to be explored from the perspective of neuroscience (e.g., Frith et al. 1991; Baumeister et al. 2009; Soon et al. 2009; Sinnott-Armstrong and Nadel 2010; Gazzaniga 2011; Haggard 2011), which clearly indicates that the brain is central to identifying and understanding the something that produces will. Any neuroscientific claim to have found the something missing from MPW should satisfy two conditions: (1) identify the I-space structures that underpin action and (2) reveal the I-space structures that operate to make action purposive. Largely based upon the findings of neuroimaging studies in what follows I propose an account of how the brain is connected with E-space in a reflexive manner that makes action possible, and additionally, I advance the idea that the brain reflects upon E-space in a cultural manner giving action its purposive direction.

The Reflexive Brain and Action

There are numerous approaches to reflexivity. From a CRS perspective, reflexivity is a fundamental attribute of *some* structures, fundamental because it is

a determinant of their powers. Power is any effect(s) of exercises of force, with force understood not narrowly as physical coercion, but broadly as any combination of resources with the ability to cause effects (Reyna 2003). Reflexive structures to varying degrees know "what is" and use this knowledge as a force resource to know "what to do about it." Reflexivity is about power because it is about structures interrogating antecedent E- and I-space events in the present to do something affecting subsequent E- and I-space actions. In this optic, reflexivity is a force resource with the power of binding past, present, and future. How does reflexivity guide action?

Reflexivity, according to Anthony Giddens, "... is generally defined as the monitored character of the ongoing flow of social life" (1984, p. 3). The key term in this definition is "monitored." In CSR, social life is monitored if there is a *monitor*, in the sense of "a device ... for observing or recording the operation of a machine or system..." (Random House Dictionary 1967, p. 925). The "system" monitored is antecedent E- and I-space events. The "observing" is representation of the states of, and plans for, these realities. Observing involves two sorts of representations of E- and I-space events: those describing their states and those involving procedures of what to do about them. In order for individuals to be reflexive, two conditions must be satisfied in their I-space. First, they need an I-space monitoring device able to represent E- and I-space events and to plan what to do about them. Second, they need two "reflexivity extension cords," which – like extension cords – plug individuals' monitoring devices into E-space. The first set of such cords brings antecedent E-space information from reality to I-space to the monitor, in effect plugging E-space into I-space. The second chord brings information from the monitor to make actions that participate in subsequent E-space events. Next it is shown how the human brain working through the nervous system comes complete with a monitoring device and two extension cords.

Wiring, Information, and Representation

The brain is a system of specialized cells called nerves or neurons. There is an estimate of 86 billion neurons (plus at least the same number of support or glial cells) (Azevedo et al. 2009). This system of densely interconnected cells is often called the brain's "wiring." It brings information about E-space into the body, monitors it, and sends it back out as action into E-space. The wiring, then, is both the extension cords and the monitoring device of human reflexivity. It is therefore important to explore this device more fully.

Information in the brain corresponds to electrochemical impulses traveling between neurons. Afferent nerves (those carrying impulses toward the brain from sensory organs) are the extension cord that plugs the brain into antecedent E-space events. Efferent nerves (carrying impulses away from the brain toward the muscles) are the extension cord that plugs the brain into subsequent E-space events through movement. Electrochemical flows within the brain correspond with the operation of the monitoring device that surveys what is out there in antecedent E-space and plan

what to do about it in subsequent E-space. These impulses flow within organizations of neurons variously termed modules, circuits, networks, pathways, or systems. Electrochemical energy moves along all of this wiring in feedback, feedforward, parallel, lateral, convergent, divergent, and hierarchical manners. Systems are collections of interconnected neurons in particular places in the brain performing various functions. They interact with other systems accomplishing higher order functions. For example, the limbic system, primarily concerned with emotions, is in turn connected with the prefrontal cortex, which is in charge of executive control. Together, these two systems form the basis for decision making.

Another important concept to introduce at this point is that of *representation*. Representation in CRS is literally the material reproduction, i.e., *re*-presentation, of something as something else. Our interest is in neuronal representation [representation, neuronal]. This is the representation in neurons of information, i.e., re-presentations of events in an actors' E-space, and what to do about them, in the past, present, and future. Past representation is memory. Present representation can be both present perception and past memories recalled in the now. Future representation is wishes about what to do. Much of the nature of neural representation remains to be learned. However, it is clear that information is represented in the brain as patterns of electrochemical impulses, "…sets of firings in a specific collection of neurons" (Baron 1987, p. 28). This "patterns" are specified by "the rate of firing of each neuron in the collection of neurons bearing the information" (Baron 1987, p. 29). Neurophilosopher Paul Churchland has called these patterns "configurations of synaptic connections," which he believes allow the "general and lasting features of the external world" to be "represented in the brain" (1995, p. 6). Consciously or unconsciously, the patterns of synaptic configurations represent E-space in I-space, either as information about current perceptions, past memories, or future wishes. In this optic, information and representation are not disembodied mental peregrinations. They are the workings of material, neuronal structures. It is time to present how the neuronal flow of information makes the brain a reflexive structure.

The Tripartite Brain

Edward Hundert (1989) proposed an influential tripartite model for the functional organization of the brain. The model divides "the brain into input, central, and output systems" (Hundert 1989, p. 201). The input system performs the tasks of detecting and perceiving. Detection (sensation) and perception provide information about "what is out there" in E-space or I-space to the brain. A man waves and says "hello" from across the street. The waving is transformed into radiant energy, involving light waves, which you can see. The "hello" is transformed into acoustical energy, involving sound waves, which you can hear. The properties of these energies – their amplitudes and frequencies – bear information about the happenings – waving and saying "hello." Energies that contact receptors are termed "stimuli," in the sense that they stimulate the receptors. These respond to stimuli from outside or inside the body and work as transducers, as they transform one form

of energy into another. "Detection" is therefore the transducing of energy bearing information from E-space or parts of I-space other than the brain into electrochemical energy in neuronal I-space, and its transmission along "...afferent pathways to the corresponding ...sensory areas" (Kolb and Wishaw 1998, p. 223).

Information flows through lower and higher afferent pathways to two sorts of sensory areas, one located largely in the posterior (also called the sensory) cortex and the other in the sub-cortical limbic areas. Consequently, two basic types of representation occur: sensory and emotional. Information relayed to the posterior cortices (parietal, temporal, and occipital) is re-presented as perception of sights, sounds, smells, etc. Information flowing to the limbic system (including anterior thalamic nuclei, hippocampus, amygdala, septum, and cingulate cortex, among others) is encoded as emotion. Sensory and emotional information is then projected to the insula, which integrates information on the physiological circumstances of the entire body and from this appears to generate feelings. Current studies suggest the insula to be "...at the very essence of human feelings" (Craig 2004, p. 239). In sum, the input system of the brain first detects and then re-presents E-space as sensory and emotional information. These representations may occur at roughly the same time and, from the actor's perspective, may seem to be reality, with it understood that this "reality" is not reality itself but representations of it. Otherwise put, the input system of the brain tells a person what is the likely state of the world (out there, but also their inner world).

Next, consider the central system where "...the meaning of information" is established (Hundert 1989, p. 201). It is located in the anterior part of the cortex, the frontal lobe, and receives perceptual and emotional information from pathways in sub-cortical regions and the sensory cortex. The most important part of the central system, just behind the forehead, is the prefrontal cortex (PFC), a complex set of areas integrating sensory and emotive representations in order to accomplish "executive control," guidance of thought, and action. The output of the PFC is a new type of information, termed "action messages," specifying what to do about what is that flows along efferent circuits to the posterior part of the frontal lobe, where the output system is located.

Output areas are those that, upon receipt of central system information, perform the task of putting the body into action. They are the premotor and the primary motor cortices. Efferent pathways from the latter transmit action messages to the basal ganglia and the cerebellum from whence they are sent through the spinal cord to the muscles. Two sorts of actions are characteristically performed. The first is practical, where the body's muscles contribute to some human practices in E-space. The second is discursive, where the body's muscles contribute some meaningful message in E-space. Consequently, it can be said that the tripartite brain is one where information from antecedent E-space and non-brain I-space arrives via the input system to the brain's I-space where representations are made. Then, in the central system, meanings of these are interpreted and, on the basis of this, are made into discursive or practical action that reenters subsequent E-space via the output system.

Earlier, discussion suggested that a structure was reflexive if: (1) it had a monitor capable of representing E- and I-space events and doing something about them

and (2) there were incoming and outgoing “extension cords”; the former plugging E- and I-space into the monitor, the later plugging the monitor back into E- and I-space. It has been shown how afferent circuits from the sense receptors to the thalamus, the limbic system, and the posterior cortex transmit information from E- and non-brain I-space events, constituting an incoming extension cord. Then, the flow of this information to the posterior cortex and limbic system gives rise to perceptual and emotive representation of these events, in effect signaling “what is” out there in E-space and non-brain I-space. Next, the flow of these representations to the PFC leads to further flow of information within the various sub-structures of the PFC, and other brain regions to which it is connected, which interpret the procedures concerning “what to do” about these representations, i.e., signaling what to do about represented E- and I-space. Thus, together, the posterior cortex, the limbic system, and the PFC operate as a monitoring device. Finally, the efferent circuits leading from the PFC to the premotor regions, to the primary motor cortex, to the cerebellum, the spinal cord, and onto the somatic motor system effect the implementation of what has been planned in the brain’s monitoring structures, and may be said to be the outgoing extension cord. In sum, the human nervous system contains the monitoring and extension cord structures necessary for reflexivity [reflexivity, and the brain]; and it is in this sense that the brain is a reflexive organ making possible action, satisfying the first of the two conditions necessary for understanding it as “the something” of human will.

One observation might be made at this point: As the brain works through flows of electrochemical energy in neuronal pathways, it is literally more accurate to say that people go with the flow than that they make decisions (rational or otherwise). This raises the question: How does the flow impart purpose to human action? This is the topic of the following section.

The Cultural Brain and Purpose

French phenomenologist Maurice Merleau-Ponty recognized that human action was “intentional,” i.e., purposive, because it “is governed by certain pre-established nervous pathways such that” the goal intended is obtained (1942, p. 7). This section argues that these pathways are cultural ones. The first task of the section is to show how culture has a double structural reality: On the one hand, it is an aspect of social system; on the other, it is embedded within the brain’s neural wiring. The second task is to show how these pathways impart direction to action. Culture and hermeneutics are introduced next; after this, it is hypothesized how the PFC uses culture to impart purposive bias to action.

Culture’s Two Realities

Hermeneutics is the study of understanding (*verstehen*), which comes about as a result of interpretation. One axiom of hermeneutics is: As people interpret,

so they act. The view that the brain is involved in interpretation is old. Baruch Spinoza in the seventeenth century remarked, "men judge things according to the organization of the brain" (in Reyna 2002, p. 96). Emmanuel Swedenborg in the eighteenth century insisted, "…the cortical substance imparts … understanding" (Reyna 2002, p. 89). What reason is there to believe that the brain "imparts" understanding? I argue next this is because the brain is a cultural neurohermeneutic system. Consideration of culture is necessary to make this argument. At its broadest, culture may be defined as that which imparts understanding to actors. But culture may be said to have two distinct manifestations: cognitive and emotive.

Cognitive culture [culture, cognitive] involves understanding of learned and shared "…semantic domains organized around numerous features of meaning…" (Tyler 1969, p. 359). "Semantic domains" are systems of linguistic symbols (words) that constitute informational categories (hereafter simply categories) and are the building blocks of messages. For example, there is the semantic domain of "animal." "Cats" and "dogs" are smaller categories within the larger domain of "animals." Arranging different symbols in grammatically appropriate manners creates a cultural message. Therefore, cognitive culture is, in part, the universe of an actor's informational categories drawn from the universe of categories available in social E-space. Interpretation in this optic is assigning of a cultural category or message to perceptions. When a mother points to a furry, meowing beast and says to her daughter "pussy," her discursive action involves "enculturation," the embedding of discursive culture in another person's neuronal wiring. When a 3-year old then sees a furry creature that meows and exclaims – "pussy" – she has reached understanding and is making an interpretation. Cognitive culture also incorporates other types knowledge present in E-space comprising episodic knowledge, that is, knowledge about events and facts, and procedural knowledge, or knowledge about how to perform a task. These different types of cognitive culture – semantic, episodic, procedural – are encoded as long-term memories in brain regions distributed across temporal, parietal, and occipital cortices as well as in sub-cortical structures (chiefly, hippocampus, amygdala, basal ganglia and cerebellum) (Squire 1987; Fuster 1995).

Cognitive culture comprises both categorizations of "what is" and procedures, or "rules," for dealing with it. Elsewhere (Reyna 2002) categorizations of "what is" have been called "perceptual," because they classify reality, and abstractions about reality, into different symbols. Instructions for dealing with different perceptual cultural categories were said to be "procedural." Cultural systems, thus, have perceptual and procedural aspects. For example, among the Barma of Chad, there is the cultural symbol *bob*. Perceptually, fathers and fathers' brothers are classified as *bob*. Procedurally *bob* are shown *hormo* (respect). Consequently, at an antecedent time, E-space$_1$, a paternal uncle arrives at his niece's hut. He is seen by her and interpreted in her I-space as a *bob*, and at a subsequent time in E-space$_2$, she bows to him, offers water, and offers food, performing the actions consistent with the procedural memory of giving *hormo*.

In addition to perceptual and cognitive, cultural representations can also have an emotive character. Two important properties of emotional memories

are valence and arousal. "Valence" is the dimension of emotion that varies from positive to negative (attractive/pleasant to aversive/unpleasant); while "arousal" is the dimension that varies from no response to intense response (calm to excited) (Schacter 2011). Emotional memories are encoded by an array of structures including the amygdala and other limbic structures like the nucleus accumbens, insula, and orbitofrontal and cingulate cortices (Phelps 2004). Such memories are an actor's emotional culture [culture, emotional], so long as these memories are acquired from and contribute to social E-space.

Cognitive and emotive cultures are closely associated. This is because the cognitive and memory systems "interact in subtle but important ways" so that they "act in concert when emotion meets memory" (Phelps 2004, p. 198). It is far from certain exactly how this occurs, but one way appears to be that the brain remembers reward together with cognitive cultural categories, and cognitive categories are also tagged by arousal and valence. This means that if a cognitive procedural cultural category is retrieved from memory, its emotive cultural representation is also likely to be retrieved. For example, if one is in the cultural category "Republican" in American politics, it has been demonstrated that the opposing cognitive cultural category, "Democrat," is emotionally remembered as something that is pretty unrewarding (Weston 2007). Consequently, when Republicans are obliged to contemplate Democrats, they indulge in "motivated reasoning" involving the operation of limbic structures (including ventromedial PFC, anterior cingulate, insular, and lateral orbital cortices (Weston et al. 2006)) that reflect the negative valence of their memories of Democrats. Dopamine appears to play a role in the association of cognitive and emotive culture by signaling reward: It is released prior to events remembered as pleasant, but absent prior to those remembered as not pleasant (Schultz 2002; Frith 2007). Thus, when a Republican sees a Democrat, there is no dopamine signal with its promise of reward.

Given the preceding, cognitive and emotional cultures might be thought of as a laminate, bonded into a common system of understanding that identifies what is, how it feels, and what to do about it. Returning to the example of the Barma, at an antecedent time, E-space$_1$, a father arrives in a daughter's hut; his daughter sees him. She interprets her father in terms of her cognitive culture as a *bob*. One's action toward a *bob* should follow a particular cultural procedure, that of giving him respect. It can be speculated that the sight of *bob* may trigger a spurt of dopamine compelling the daughter into action. Limbic structures may be variously activated, producing at a subsequent time in E-space$_2$ the performance of *hormo*, leaving her feeling good about it. Here, then, is the significance of culture. It allows the brain to achieve understanding: cognitively of what is and what to do about it; while emotively motivating action consistent with cognitive understanding. Consequently, it is culture that imparts direction to action, and it does so, not because people make decisions, but because they go with the electrochemical flows of their neuronal culture.

The Prefrontal Cortex and Cultural Purpose

A way of beginning to answer how people "go with the flow of their culture" as it is encoded in the brain is to reflect upon the structure of neural wiring. There is an efferent, convergent hierarchy of neurons from the posterior cortex and the limbic system, where cognitive and emotional representations of E-space events are achieved, that terminates in the PFC, and an afferent, divergent hierarchy that leads from the PFC to the motor cortex, the spinal cord, the muscles, and from thence back out into E-space events as action. Something happens within the PFC that "processes" what is going on in E-space and responds by creating action responses, suggesting the PFC is the place in the brain where will to act is generated. Of course, the next query is, how does the PFC "processes" incoming electrochemical flows to make them into outgoing electrochemical flows? At least eight theories had been proposed at the turn of the century to account for PFC function (Wood and Grafman 2003), but none of these specifically addressed the role of culture in PFC function. What follows is a speculative account of how the PFC might employ the culture that it has embedded in its neural tissues to produce action. PFC [prefrontal cortex, structure, and function] anatomy is described first, including its parts and connections with other parts of the brain, and their joint functioning. Next, this information is analyzed to suggest how the PFC uses culture to make action.

The PFC is the foremost part of the frontal lobe. On its sides are the dorsolateral PFC (DLPFC) beneath which is the ventrolateral PFC (VLPFC). More ventrally and medially, behind the eyes, is the orbitofrontal PFC (OPFC). Finally, the ventromedial PFC (VMPFC) is located along the midline and above the OPFC. The different PFC regions are connected with each other. There appear to be "*two parallel*" neural circuits within the PFC (Pochon et al. 2002, p. 5669). The first circuit involves the DLPFC and the VLPFC. It is concerned with *planning* and *executive control of action* (Pochon et al. 2002, p. 5669). The second circuit includes the OPFC and the VMPFC, which deal with emotion and *motivation* (Pochon et al. 2002, p. 5669). The parallel circuits, however, are themselves connected by feedback loops (Fuster 2008). The PFC is as well richly connected with other brain areas. There are also dorsal (upper) and ventral (lower) pathways leading from the posterior cortex to the DLPFC and VLPFC, respectively (O'Reilly 2010). The ventral pathway, largely underpinned by the temporal cortex, has been termed the "what" circuit as it *extracts information relevant for identification and other forms of semantic knowledge*; the dorsal pathway, on the other hand, largely underpinned by the parietal cortex, has been termed the "how" circuit, because . . .*it extracts visual signals relevant for driving motor behavior* (perception for action) (O'Reilly 2010, p. 355, emphasis in the original).

While much debate exists surrounding functional segregation in the PFC (Badre 2008; Botvinick 2008), enough is known to sketch its functional organization. There is agreement that the DLPFC functions as the brain's "working memory" system (Baddeley 1986), which broadly corresponds to conscious thinking about

something. In fact, for some neuroscientists, consciousness itself is understood as "…awareness of what is in working memory" (LeDoux 1996, p. 278). The "work" in working memory, according to Carter, refers to the DLPFC taking "… information from many different parts of the brain–sensory and conceptual– …" which it "'juggles' to come up with ideas for action" (2002, p. 235).

The VLPFC appears involved with language and the retrieval and maintenance of visual and or spatial information. It is thought to be the place where perceptions are put into words. There is also considerable research indicating that it is the place where emotional regulation takes place (Cohen et al. 2012). The VLPFC works closely with the DLPFC. Ward (2010) argues that the VLPFC integrates semantic and perceptual information, stored largely in the posterior cortices, together with emotional information from the limbic system, and that the DLPFC manipulates this new informational composite.

Functional segregation within the PFC is also illustrated by research into inductive and deductive inference. Monti et al. (2007) has made a distinction between "core" and support deduction areas. Both appear "primarily in the left hemisphere" (2006, p. 32). The "core" performs "deductive operations," while the "support" area maintains "content-specific …semantic information" (2006, p. 2). The "core" area appears to be located in the DLPFC, whereas the "support" area seems to be positioned in the VLPFC – with the support area providing the semantic information about which the core area makes deductive inferences. Regarding induction, Goel et al. found that it activates "…a large area comprised of the left medial frontal gyrus, the left cingulate gyrus, and the left superior frontal gyrus" (1997, p. 1305).

The OPFC and VMPFC are emotion and reward centers of the brain (Rolls 2005). As Carter puts it, *Emotions become conscious here*, because it is the place "…where emotions are experienced and meaning bestowed on our perceptions" (2002, pp. 115, 210). Specifically, the OPFC appears to do this by "…decoding reinforcement contingencies in the environment" (Rolls 2005, p. 140). According to LeDoux, people "with orbital cortex damage become oblivious to social and emotional cues, have poor decision-making abilities, and some exhibit sociopathic behavior" (2002, p. 227). This is likely to be the case, because with a damaged OPFC, such persons have lost their knowledge of the reinforcement contingencies in their E-spaces. They are literally cast adrift in a world in which they have been disconnected from what is emotionally beneficial and disadvantageous out there. Their life is, to appropriate a term from Emile Durkheim, neurological anomie. Durkheim's "anomie," articulated in *Suicide* (1898 [1966]) was "normless," not having any rules to know what to do. "Neurological anomie" is the loss of knowledge of the contingencies of reinforcement of everyday life, and so not knowing the emotional rules of what to do. Such life is not so much sociopathic as it is tragic: because persons enduring it are (emotionally) lost in E-space.

Two other brain structures that play an important support role to PFC function are the insula and the anterior cingulate cortex (ACC). A structure where inputs converge from all across the body, the insula, as noted earlier, integrates sensory

and emotional information to generate a representation of the physiological circumstances of the entire body from which a sense of bodily self-awareness (Craig 2009) as well as a sense of agency (Farrer and Frith 2002) emerges. The ACC, on the other hand, appears to have a particularly wide range of functions, including heart rate and blood pressure regulation. Studies of this region suggest that it plays both a "prominent role in executive control of cognition" (Carter et al. 1998, pp. 747–48) and in *emotion/feeling* (Damasio 1994, p. 71). It also "helps to focus attention" (Carter 2002, p. 210), is active in error detection (Carter et al. 1998; Bush et al. 2000), assists conflict monitoring (Holyrod et al. 2004), and is involved with reward-based learning. Attention, error detection, conflict monitoring, and reward learning are interrelated functions. The function of the ACC is therefore likely to detect and monitor actions to respond to errors on the basis of remembered contingencies of reinforcement. ACC and DLPFC operate very close together. Both regions tend to be activated at the same time. However, recent research suggests that the dorsolateral prefrontal cortex functions to "implement control" over action, while the anterior cingulate cortex works in the "monitoring" of the "performance" of action (McDonald et al. 2000, p. 1835).

In sum, neuroscience has revealed that the different parts of the PFC function to provide executive control by processing sensory together with emotional and proprioceptive inputs both from E- and I-space. This processing involves provision of working memory, planning, monitoring, inhibition, switching of attention, and cognitive flexibility. Finally, though much remains uncertain, these operations are about what people conventionally understand as thinking and feeling what is good to do (Fuster 2008).

Where is culture in PFC [prefrontal cortex, and culture] functioning? We know that the PFC plays a central role in people's actions. PFC function underpins feeling and thinking about what is good to do. Culture provides individuals guidelines about what is "good" (and what is not). I would propose culture does this by biasing PFC function in the following manner: To act on a bias is to act in a certain direction, which is another way of saying to act with purpose. Consider that certain events occur in E-space. Information about these events makes its way from E-space to the PFC (through receptor organs, posterior cortices, and the limbic system). The ventral and medial aspects of the PFC (OPFC/VMPFC) determine the reward contingencies of the emotional representations. This information is then transmitted to the lateral PFC pathway where emotional representations become associated with different cognitive, perceptual, and procedural representations. A key point is that both emotional and cognitive representations are cultural: They are acquired from the surrounding cultural milieu. Emotive and cognitive cultural representations are therefore able to bias DLPFC activity whereby on-line (i.e., within working memory) inferences can be made regarding "what is good to do." (All this is accomplished with the support from the insula and the ACC: the former by integrating sensory and emotional information making it available to consciousness, and the latter by screening activity to determine if it follows the action bias established by the DLPFC. If it does not, the ACC signals the DLPFC to re-establish the action bias).

The PFC [prefrontal cortex, as cultural neurohermeneutic system] establishes bias, because it is a cultural neurohermeneutic system, that is, it receives sensory, proprioceptive, and emotional information from E- and I-space, and operates on it based on previous cultural representations across these domains to discover “what is” and “what to do about it.” Such a discovery is an interpretation and, on the basis of such interpretation, the PFC activates the premotor and motor cortices that, in turn, generate action. The implications of the foregoing discussion are threefold: First, the human brain understood as a monitoring device means that it is able to reflect upon events in reality; second, the PFC understood as a biasing device means that it is able to set direction for action giving the brain power to impart purpose on reflection; finally, PFC bias understood as cultural means that purpose is itself cultural.

Free Will and Agency

There are two important implications from the account thus far: The first concerns the modern problematic of will (MPW); the second the very possibility of free will, which leads to a re-appraisal of agency.

A problem with the MPW is the lack of clarity about the something that performs the functions of will, so well articulated by MPW philosophers. The preceding analysis has suggested that a culturally reflexive brain performs the functions of will. Remember that these functions comprise, first, producing action and, second, manifesting purpose, which can be done by rational or non-rational means. It is the brain’s reflexivity that has been shown in the present account to fulfill the function of producing action. In this context, the cultural neurohermeneutic system in the PFC has been proposed to fulfill the function of imparting purpose to action. Finally, the suggestion that the DLPFC/VLPFC and the OPFC/VMPFC establish purpose with both cognitive (i.e., more rational) and emotive representations of events suggests that both sets of MPW philosophers were correct. Will, it seems, operates through both rational and non-rational means.

Consider, however, the implications of the culturally reflexive brain for the possibility of free will [free will, possibility of]. Here I believe many thinkers have been barking up the wrong tree. They want to know: Is there, or is there not, free will? Certainly, prior to any action, the biasing device, that is, the cultural neurohermeneutic system, is *already there* in the PFC (as a result of enculturation), and what is willed is a consequence of the operation of this system. If this is the case, and if freedom is about “unrestricted” action, then, the notion of free will is implausible, because willed action is always restricted by the properties of the brain. The counter argument, “But I am free to change my mind” only means that one set of electrochemical flows in the brain operates, not another, and it is those electrochemical flows which determine will. The retort, “But I am free to change my mind to change my electrochemical flows,” ignores that fact that it is either other electrochemical flows in your brain that change your mind, or events in E-space that change the electrochemical flows in your brain. The sense that you

are free to choose is a phantasmagoric (mis)representation of the reality that you go with the flow of electrochemical messages in your brain.

Furthermore, from a critical perspective, belief in free will [free will, critique of] can assist the powerful in their domination and is actually of little practical interest when seeking to help people in their everyday lives. Consider, first, how the construct of free will assists elite domination. If one believes in free will, it is possible to classify people as “immoral” or “criminal” when they do something that breaks an ethical or moral rule, because they would have been free to will their actions differently. However, many recognize that ethical and moral canons often benefit the powerful. Many poor people steal because of something in their E-space. They are poor. Rich folk classify thieves and the like as “the dangerous sort,” because they – in this example, the poor who steal – have free will and could have avoided a life of crime, thereby avoiding their own immorality – that of tolerating and/or causing poverty – by classifying the poor as immoral criminals.

Let us explore whether the notion of free will has some practical utility when seeking to help people in their everyday lives. It was earlier established that there were forces that influenced will. These were restrictors and expediters that were both internal and external to the human body. This is to say that there were either E- or I-space events that helped humans achieve their will or hampered its achievement. Current structures globally dominating E-space – those of capitalism and imperialism – organize both E- and I-space into two classes – one containing very few actors with enormous external resources to expedite their wills, and the other one containing most of humanity with very few external resources to expedite their wills. This is underpinned by a set of circumstances whereby what those in power do to realize their will often restricts powerless actors’ attempts to realize, in turn, their own wills.

It is at this juncture that the notion of agency becomes relevant. There is a large literature on agency (for a review, see Emirbayer and Mische 1998). The approach suggested by Ortner (2006) is useful for the present argument. For Ortner, agency “is about power” and “the pursuit of (culturally defined) projects” (2006, p. 139) (that is, projects exhibiting will). Thus, “agency” is the power to realize will. Consequently, in the CSR optic, what is important is first to distinguish worthwhile from worthless will, and then to devise methods of expediting the agency of actors striving for worthwhile wills and restricting it for the agency of actors striving for worthless wills. The theoretical notion of a “worthwhile” will [will, worthwhile] is contentious. Practically speaking, however, it seems worthwhile that people have decent places to live, nourishing food, education, and healthcare. The hope is that by strengthening ordinary peoples’ agency, they are given the resources to achieve the power of realizing more of their wills.

Conclusion: By Way of a Eureka Moment

For millennium upon millennium, the search for free will went on, as if it was the Holy Grail of human inquiry. Eureka, the Holy Grail was a phantasmagoric,

and so is free will! This chapter has advanced the idea that the will is something, and that something cannot come from nothing, i.e., that something that is will is either determined within the cultural, reflexive brain in I-space or in the events of social and natural systems in E-space that connect with the brain's I-space. This being the case it is pragmatically argued that it makes critical sense to expand ordinary peoples' agency, the better to give them the power to realize worthwhile wills.

References

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., et al. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology, 513*(5), 532–541.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences, 12*, 193–200.

Baron, R. (1987). *The cerebral computer: An introduction to the computational structure of the human brain*. Hillsdale: Erlbaum.

Baumeister, R., Crescioni, A. W., & Alquist, J. (2009). Free will as advanced action control for human social life and culture. *Neuroethics, 4*, 1–11.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences, 12*, 201–208.

Carter, R. (1998). *Mapping the mind*. Berkeley: University of California Press.

Carter, R. (2002). *Exploring consciousness*. Berkeley: University of California Press.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection and the on-line monitoring of performance. *Science, 280*, 743–749.

Churchland, P. (1995). *The engine of reason, the seat of the soul: A philosophic journey to the brain*. Cambridge, MA: The MIT Press.

Cohen, J., Berkman, E. T., & Lieberman, M. D. (2012). Intentional and incidental self-control in ventrolateral prefrontal cortex. In D. Stuss & R. Knight (Eds.), *Principles of frontal lobe function*. New York: Oxford University Press.

Craig, A. D. (2004). Human feelings: Why some are more aware than others. *Trends in Cognitive Science, 8*(6), 239–241.

Craig, A. D. (2009). How do you feel now? The anterior insula and human awareness. *Nature Reviews. Neuroscience, 10*(1), 59–70.

Damasio, A. R. (1994). *Descartes' error*. New York: Avon Books.

Durkheim, E. (1898 [1966]). *Suicide*. New York: Macmillan.

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology, 103*(4), 962–1023.

Farrer, C., & Frith, C. (2002). Experiencing oneself as another person as being the cause of action: The neural correlates of the experience of agency. *NeuroImage, 15*(3), 596–603.

Freud, S. (1935). *A general introduction to psychoanalysis: A course of twenty-eight lectures delivered at the University of Vienna*. New York: Liveright.

Freud, S. (1950). In N. Fodor & F. Gaynor (Eds.), *Freud: A dictionary of psychoanalysis*. New York: Philosophical Library.

Frith, C. (2007). *Making up the mind: How the brain creates our mental world*. Malden: Blackwell.

Frith, C., Friston, K., Liddle, P., & Frackowiak, R. (1991). Willed action and the prefrontal cortex in man: A study with PET. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 244*, 241–246.

Fuster, J. M. (1995). *Memory in the cerebral cortex*. Cambridge, MA: MIT Press.

Fuster, J. M. (2008). *The prefrontal cortex*. Amsterdam: Academic.

Gazzaniga, M. S. (2011). *Who's in charge? Free will and the science of the brain*. New York: Ecco.

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Los Angeles: University of California Press.

Goel, V., Gold, B., Kapur, S., & Houle, S. (1997). The seats of reason? An imaging study of deductive and inductive reasoning. *Neuroreport, 8*(5), 1305–10.

Haggard, P. (2011). Decision time for free will. *Neuron, 69*(3), 404–406.

Horkheimer, M. (1982). *Critical theory*. New York: Seabury Press.

Hume, D. (1729–1740 [1978]). *A treatise of human nature*. New York: Oxford.

Hundert, E. (1989). *Philosophy, psychiatry and neuroscience*. New York: Oxford University Press.

Kant, I. (1785 [1997]). *The groundwork of the metaphysics of morals*. Cambridge: University of Cambridge Press.

Kant, I. (1788 [1956]). *Critique of practical reason*. New York: Bobbs-Merrill.

Kolb, B., & Wishaw, I. (1998). Brain plasticity and behavior. *Annual Reviews of Psychology, 49*, 43–64.

LeDoux, J. (1996). *The emotional brain: The mysterious underpinning of emotional life*. New York: Touchstone Books.

LeDoux, J. (2002). *Synaptic self: How our brains become who we are*. New York: Penguin.

Merleau-Ponty, M. (1942). *La structure du comportement*. Paris: PUF.

Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: A language-independent distributed network. *NeuroImage, 37*(3), 1005–1016.

Nietzsche, F. (1886 [1989]). *Beyond good and evil: Prelude to a philosophy of the future* (trans: Kaufmann, W.). New York: Vintage.

Nietzsche, F. (1901 [1968]). *The will to power* (trans: Kaufmann, W., & Hollingdale, R. J.). New York: Vintage.

O'Reilly, R. (2010). The what and how of prefrontal cortical organization. *Trends in Neurosciences, 33*(8), 355–361.

Ortner, S. (2006). *Anthropology and social theory*. Durham: Duke University Press.

Oxford English Dictionary. (1989). *The Oxford English dictionary* (Vol. VI). Oxford: Oxford University Press.

Pochon, J. B., Levy, R., Fossati, P., Lehericy, S., Poline, J. P., Bihan, D. L., & Dubois, B. (2002). The neural system that bridges reward and cognition in humans: A fMRI study. *Proceedings of the National Academy of Science, 99*, 5669–5674.

Reyna, S. (2002). *Connections: Brain, mind and culture in a Social Anthropology*. London: Routledge.

Reyna, S. (2003). Force, power, and the problem of order: An anthropological approach. *Sociologus, 3*(2).

Reyna, S. (2012). Neo-boasianism, a form of critical structural realism: It's better than the alternative. *Anthropological Theory, 12*(1), 73–101.

Rolls, E. (2005). *Emotion explained*. New York: Oxford University Press.

Schacter, D. L. (2011). *Psychology* (2nd ed.). New York: Worth.

Schopenhauer, A. (2000 [1818]). *The world as will and representation*. New York: Dover Editions.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36*(2), 241–263.

Sinnott-Armstrong, W., & Nadel, L. (2010). *Conscious will and responsibility: A tribute of Benjamin Libet*. Oxford: Oxford University Press.

Soon, C., Brass, M., Heinze, H., & Haynes, J. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience, 11*(5), 543–545.

Ward, J. (2010). *A student's guide to cognitive neuroscience* (2nd ed.). London: Psychology Press.

Weston, D. (2007). *The political brain: The role of emotion in deciding the fate of the nation*. New York: Public Affairs.

Weston, D., Blagov, P., Harenski, K., Kilts, C., & Hamm, S. (2006). Neural basis of motivated reason: A fMRI study of emotional constraints on partisan political judgments in the 2004 U.S. presidential election. *Journal of Cognitive Neuroscience, 18*(11), 1947–1958.

Wood, J., & Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews. Neuroscience, 4*, 139–147.

What Is Normal? A Historical Survey and Neuroanthropological Perspective

21

Paul H. Mason

Contents

Abstract

What is considered typical and usual is guided by the cultural framework a person is accustomed to. In the brain sciences, it can easily be forgotten that "normal" and "normality" are not rock solid concepts. Simply acknowledging that "normal" does not have an objective existence is insufficient without also changing scientific practices accordingly. This chapter unpacks why normality has been such a persistent concept for the last two centuries. The concept of normality grew alongside the development of statistical methods and was instrumental in constructing a much maligned concept of "degeneration." Statistics are useful in a wide range of scientific contexts, but detrimental when used as a blunt instrument of measurement to legitimize labels that differentially sort people into subpopulations that augment social inequalities. A rigorous questioning of normality and degeneration ensures an ethical engagement with hypotheses of neuroscience experiments and the implications of research findings. This chapter surveys some of the key historical developments at the origins of the brain sciences to understand some of the biases present today. The language used to classify the world can lead to blind spots that remain

P.H. Mason
Woolcock Institute of Medical Research, Glebe, NSW, Australia
e-mail: paul.mason@woolcock.org.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_153,

hidden for generations. Rather than searching for a direct localization of human behavior in biological etiology, this chapter advocates a complex localization through mapping distributed agency across intersecting neurobiological, cultural, and environmental processes. Normal might be a value-laden term that has no place in the brain sciences, but a value-free operational conceptualization of the processes of degeneracy may be central to understanding dynamic neuro-cultural systems.

Introduction

Normality is a term that recurs with disturbing frequency in the writings of psychologists, psychiatrists, psychoanalysts, sociologists, and other people concerned with human behaviour.
H.J. Eysenck 1953, p. 177

What is a normal human being?
John Blacking 1977, p. 12

The normal is what you find but rarely. The normal is an ideal. It is a picture that one fabricates of the average characteristics of men, and to find them all in a single man is hardly to be expected.
Somerset Maugham 1938, p. 67

In a telling twentieth century investigation (Rosenhan 1973), eight pseudopatients presented themselves with false symptoms to twelve different medical institutions. Each pseudopatient falsified existential auditory hallucinations to medical staff in order to be admitted into a psychiatric ward. Their falsified self-reports were similar to known psychiatric conditions but were absent from the medical literature of the time. Upon admission, the pseudopatients did not report further hallucinations and behaved sanely and cooperatively. As inpatients, they took notes of their observations and experiences. During their time in various psychiatric wards, they observed that a patient's diagnosis influenced the way staff interpreted behaviors and life histories. Arriving early for a meal, for example, was interpreted as characteristic of the oral-acquisitive nature of a syndrome as opposed to the fact that besides mealtime, there were very few things to anticipate in a psychiatric hospital. For anywhere up to 52 days, despite initial false symptoms and "normal" behavior, psychiatric staff did not detect sanity among the pseudopatients. Paradoxically, other inpatients "recognized normality when staff did not" (1973, p. 252).

Rosenhan's investigation revealed how difficult "normality" can be to detect, most notably by the health professionals of his time. While we may believe that abnormal is defined by identifiable criteria, how is normal defined? Normal and abnormal are mutable concepts that have changed throughout their short history. Anthropologists understand that these terms are relative and defined by cultural experience and social background. What is abnormal in one culture may be considered normal in another. For example, being depressed may be diagnosed as a medical disorder by one authority or regarded as being a good Buddhist by another (Obeyesekere 1985); schizophrenia may be understood as a mental condition by one group or tied to shamanic practices in another (Silverman 1967); dissociative experiences may be pathologized as a neuropsychological disorder in

one community but revered as saintly trance and possession elsewhere (Seligman and Krimayer 2008); sleep paralysis may lead to Sudden Unexpected Nocturnal Death among some peoples but be completely benign among others (Adler 2011). In each case example, anthropologists have observed that a disruption in one social group is benign or healthily integrated in another. Function or dysfunction depends on context.

Normal and abnormal cannot be defined without understanding the beliefs, values, and power structures of a cultural group. Certain cultural beliefs and practices may contain protective mechanisms against some forms of mental illness but open the space for others. As the purveyors of what is normal and what is abnormal, professionals working in mental health and the brain sciences have to recognize their pivotal role in shaping societal values. Defining normality has very real implications for individuals. The opinion of a professional about a person's mental condition can lead to a range of outcomes given the situation: A diagnosis of abnormal can put someone into an institution, let them off a criminal charge, take their children away from them, give them access to support services, delegitimize their status, award them a job, impede them from a raise, distance them from other members of a community, exacerbate existing mental conditions, or even lead to suicide. In cultures where willful deviance is crime and unwillful deviance is illness (Conrad 1981, p. 107), individuals can deny moral, personal, and social controversies by feigning a psychiatric condition that works the mental health system in their favor (Szasz 1962). Political differences can also be undermined by the authoritative power claiming that the deviant position is a result of a mental condition. Medical science reserves the right to confer labels of normality and abnormality, but to what extent are these terms objective and purely descriptive?

Degenerate behavior was once considered sin during religious times, then crime during the rise of the nation state, and now a medical problem with the growing monopoly of biomedical science. Eminent scholars such as Michel Foucault (1965, 1972), Erving Goffman (1961), R.D. Laing (1960), Thomas Szasz (1962, 1963), Ivan Illich (1976), and Ian Hacking (1986, 2001) have argued that our concept of "normal" is a social construct that relies heavily on medical discourse. While the works of these scholars have been highly influential in the humanities, their impact throughout the brain sciences is negligible. Theoretical challenges to the concept of normal in psychoanalysis (e.g., Eysenck 1952) and psychiatry (e.g., Eysenck 1975) have rippled throughout these fields but have not led to tangible change (Treacher and Baruch 1981, p. 120). Interdisciplinary researchers able to translate the importance of critical perspectives from the humanities to contemporary scientific audiences will ensure that a reflexive approach to experimental research and an ethical dissemination of research findings can be carried out. Social constructionist models (e.g., Amundson 2000; Amundson and Tresky 2007; Davis 1995; Hacking 1986; Oliver 1983; Shakespeare 2010) have already proven liberating for communities of people with physical and mental handicaps, because these models demonstrate that marginalizing and stigmatizing classification schemes can be challenged, because labels of abnormal or degenerate are neither purely descriptive nor objective. A sincere engagement with the theoretical import of philosophical work on scientific practice has wide-reaching implications on the lives of individuals and the shape of societies.

H.J. Eysenck, who in the epigraph commented upon the wide imprudent usage of the word "normal," observed that "the same writer frequently employs [the word normal] now in one sense, now in another" (1953, p. 177). He found that the word "normal" was conflated with categories as vastly different as "natural," "ideal," or "statistically representative." When the word "normal" is used to refer to natural processes, Eysenck fittingly reminded his readers that normality and naturalness should not be confused, because what constitutes "natural" is relative to cultural values, beliefs, and standards (pp. 178–181). Eysenck used simple but convincing zoological and anthropological examples to demonstrate the culturally influenced interpretations of what is natural. Zoologists have revealed that what is considered instinctual, innate, and therefore natural in an animal species might be eliminated through altered developmental conditions, and anthropologists have shown that what is considered natural for human kinship relations in one context is considered unnatural in another (Ibid.). The use of the word "natural" to describe animal and human behavior is a social construction.

Surprisingly, Eysenck did not use the same social constructionist argument to counter the conflation of the word "normal" with "ideal." Eysenck's concern with conflating the word "normal" with "ideal" is based upon the claim that the "ideal" is statistically infrequent or not found in a population. While this may be a valid point, the claim that the ideal is rare or absent overlooks that "ideal" is a fabricated notion based on social values and personal attitudes. The ideals of beauty, for example, are highly variable within and across cultures. While amassing wealth and keeping slim may seem attractive in countries such as Australia and the USA, these practices have antisocial connotations in rural Jamaica where plumpness is the ideal (Sobo 1993). Ideals in one culture may be vastly different to the ideals of another. Ideals may even change over a lifetime. What is ideal is as much based on cultural values, beliefs, and standards as the notion of what is natural, and neither term should be allied with what is normal.

When it came to a statistical definition of normality, Eysenck found no issue. This was a huge oversight. Eysenck did not examine the social construction of statistics. He perceived the statistical definition of normality as "perfectly clear, straightforward, and intelligible" (p. 177). However, how can this assertion be trusted when Eysenck conflated the terms "majority" and "average" (Ibid.)? Eysenck also overlooked that any particular average characteristic within a population, such as the average height or average weight, may be exhibited by absolutely none of the individuals from that population. As the statistician Reichmann (1964, p. 121) pointed out, "The average man does not exist. The average applies to a set of data and not to an individual and if the average man could exist he would be such an odd specimen that by his very uniqueness he would deny his own title." The critical approach that Eysenck used to examine what is natural should also be deployed to understand how both the "ideal concept of normality" and the "statistical definition of normality" are socially orchestrated, historically instituted, and culturally constructed.

The walls of institutions that segregated degenerate individuals from healthy populations played an important role in creating a perception of degenerates as a clear-cut, homogeneous, and bounded population

(Adriaens and Block 2013, p. 13). The illusion of homogeneity facilitated the tendency to essentialize mental disorders despite the huge variation in the population of psychiatric patients (p. 6). Seduced by quantitative methods, systems of measurement were deployed to support these views. Privileging some forms of data over others, measurement renders some processes visible and others invisible. Measurement can be used successfully in bioassays, brain scans, and performance tasks, but it is poor at evaluating subjective experience, social variables, and cultural factors. Despite its limitations, measurement has an aura of objectivity. Measurement, however, is limited by what can be quantified. The desire for quantification and objectivity restricts the scientific gaze and is thus a concern for neuroethics.

Measurement does not just quantify the world, it also produces the world. An ethical neuroscience needs to recognize how measurement systems are a mode of power harnessed by those able to use it. Neuroscience has to be reflexive and critical to situate its biases, oversights, and unseen agendas. Anthropology is a useful addition to neuroscience, because anthropologists ask how forms of knowledge are constructed. Anthropology also offers methods to study the subjective and inter-subjective worlds of personal, social, and cultural experience. The benefits are mutual, and a better understanding of neural mechanisms from neuroscience research has an important role to play in the interpretation of cultural meanings and practices. While science makes progress by questioning research findings, anthropology advances by questioning research premises. In a move toward ethical research, the primary objective of this chapter is to unpack statistical notions of normal and historically propagated ideas of degeneration. Combining critical methods from anthropology and neuroscience, this chapter also advocates an operational conceptualization of human behavior as a constellation of distributed processes that each fractionally influence the overall function of neuro-cultural systems.

What Is Normal?

The belief that planets move around the sun in ellipses is owed to an astronomer who spent 6 years battling with a mathematical error of 8 min in the planetary motion of Mars. Many people alive today do not know who that scholar was, and yet if it were not for that scholar, then astronomy may still be imprisoned in a geocentric Ptolemaic universe. With humility, one may ask, how much of our shared reality rests upon the shoulders of people we do not know? Science is built upon the work of countless scholars who have each contributed to our collective knowledge. The history of science provides us with an important check on the historical trends that have led to current beliefs, cultural values, and scientific assumptions.

Planets do not perfectly orbit the sun. They wobble a little. To work out the movement of a planet, astronomers have to work out its location at various points and then average out the measurements. The mathematical methods of astronomy

proved appealing to nineteenth century intellectuals who were looking for new mathematical tools to study and represent human populations. Statistics was once known as political arithmetic (Davis 1995), and early statisticians borrowed concepts from astronomers such as the law of error and applied it to variable characteristics found in human populations (Vertinsky 2002). However, applying the mathematical methods of astronomy to human populations is problematical. A planet is a single moving object. A human population, on the other hand, is composed of a collection of discrete individual organisms each with their own developmental variations. Using calculations intended for a single moving object to represent populations can lead to gross misrepresentations of the data. Calculating the average motion of all the planets in our galaxy, for example, would lead to a nonsensical result. Furthermore, this nonsensical average would in no way be representative of galaxies elsewhere or indeed of all galaxies everywhere.

Despite the flaws, a belief in the regularity of statistical events led to the formulation of the average as being representative of a normal type. By giving "norms" and "normalcy" in populations a measurable character, statisticians like Adolphe Quételet (1835) and Francis Galton (1907) provided nineteenth century bourgeois middle-class values a fashionable scientific justification. The attributes of the average individual were believed to represent "all which is grand, beautiful, and excellent" (Quételet 1842). Deviations from the mean were associated with misshapenness, ugliness, and unfitness. Quetelet's use of the "normal curve" has been described as erroneous and the term "representative" as a bad slip of impartial statistical usage (Vertinsky 2002, p. 101). Although Karl Pearson is said to have corrected Quetelet's language by introducing the term "standard deviation" (Ibid.), statistics no doubt promoted the ranking of individuals within a population and promoted an ethnocentrically and politically skewed idea of what is normal. Statistics reinforced the concept of the ideal and deviant types within a population, and despite some of the most sophisticated advances in statistical measurement, certain basic presuppositions persist today.

An ethnocentric bias means that human groups tend to think of themselves as the rule and not the exception. However, contemporary statistical notions of normal are undermined by the discovery that most behavioral science theory is built upon research that examines an intensely narrow sample of human variation. In an empirical review of the top psychology journals, Henrich et al. (2010a, b) found that most studies disproportionately sample US university undergraduates who are Western, Educated, Industrialized, Rich, and Democratic (WEIRD). Henrich et al. uncovered that in the context of global populations, these WEIRD subjects tend to be outliers on a range of diverse culturally variable traits including visual perception, sense of fairness, cooperation, analytic reasoning, spatial cognition, and memory, among other basic psychological traits. Furthermore, the cultural and developmental environment for WEIRD children is highly peculiar, statistically unusual, and a distortion of species-typical ontogeny. For example, creating a separate room for a baby to sleep is an extremely uncommon practice among world societies. Henrich et al. observe that even though researchers regularly only sample from single subpopulations, they routinely assume that their results are

broadly representative and generalizable despite insufficient evidence. For a range of reasons pertaining to their unique and disproportionate particularities, WEIRD subjects may in fact be the worst population from which to make generalizations. More complete investigations with a comparison of measures across diverse populations would offer more robust representations of human variability. To construct developmental and evolutionary theories of human behavior, brain scientists should be concerned with diversity and variation rather than summarizing data using averages into one exemplar.

Statistics is an influential way of representing population characteristics that plays an influential role in normalizing human behavior. By definition, members of a population fall below the ideal. However, with the introduction of the concept of the average, the ideal was given a form that became an imperative for entire populations. This fabricated concept was rapidly adopted, because it seemingly came from authoritative sources using scientific methods. As the concept of the average spread, it fuelled one of the central forces that shaped the modern world: consumerism (Mason 2010, p. 285). The notion of the average drove consumerism, because it led people to strive for an ideal constantly out of reach. Consumption, once considered a biological disease of wasting (Sanders 1808; Epps 1859), became an accepted cultural label for a way of life.

As Western modes of consumption spread across the globe, so too did Western models of psychopathology. Watters (2010) finds a strong feedback loop between public and professional attention in the spread of Western models of psychopathology around the world. His compelling case-study of anorexia in Hong Kong (pp. 9–63) is an excellent illustration of how professionals, patients, and the public can shape and maintain, albeit unintentionally, Western models of mental illness. One potential consequence of the globalization of Western psychiatry is that it could eventually steamroll local variations in mental illness and perhaps override local buffers against pathology. Fiji, for example, had an extremely low prevalence of eating disorders before the mid-1990s, but just 3 years after the introduction of American television and a cash economy, up to 11.3 % of young girls had developed some form of eating disorder not only through restrictive dieting, but also through using purgatives and self-induced vomiting (Becker et al. 2002). American television and a cash economy therefore seem to have had a normalizing, moralizing, and stigmatizing effect on Fijian society. These forces are all drivers of homogeny, and as they act, they also open the space for the homogenizing treatments of Western neuropsychiatry.

Normalizing behavior is highly visible in social attitudes toward parenthood and child development. Caregivers are constantly subjected to a flurry of conflicting messages as to the best and most effective parenting techniques and practices. These ideal models are often based on little more than folk psychology. A cross-cultural comparison of childhood development among Samoans residing on the island of Upolu (1978–1988), the Matsigenka of the Peruvian Amazon (1995–2009), and middle-class U.S. families in Los Angeles (2001–2004) demonstrated that regardless of highly varied child-rearing practices, human development is fairly robust (Ochs and Izquierdo 2009). Child rearing varied in

terms of bodily orientation between parent and child, social attunement, levels of infant responsibility, the fostering of self-reliance, apprenticeship, and other practices. Consistent across all three groups was the display of moralizing and normalizing behavior. In contemporary society, popular media has taken this moralizing and normalizing behavior in new directions. Nowhere is parental moralizing seen more flagrantly than in product advertisements targeted at anxious caregivers. Discussions about the brain are currently a major space for moralizing discourses (Dumit 1997; Roepstoff 1999, 2002). When the moralizing of brain development becomes commoditized, neuroscience research can be co-opted for capitalist gain. From the marketing of brainfoods to the development of smart drugs, the commodification of neuroscience has ramifications for the types of research that gain financial support and the types of research that remain unfunded. Normalizing behavior may have a lot to do with power structures and group cohesion, but it should be mindfully attended to in the field studies and laboratory experiments of the brain sciences.

In brain imaging, the practice of normalization consists of transferring the scan of an individual's brain from a native to a standard space (such as Talairach or MNI space). This process allows similarities and differences across brains to be identified. However, structurally changing a brain scan to match a reference template wipes out a lot of individual variation. In transforming research subjects into research objects, the representations of each individual scan are mapped to one common co-ordinate system that allows data to be treated as if it were all recorded from the same generalized brain (Roepstorff 2002, p. 161). In other words, tremendous variability in brain activity is translated into the visual simplicity of bright, colorful, and influential brain images. Color-coding of selected statistical parameters has an iconic quality that can create the illusion of stable categories of brain function (Roepstorff 2004, p. 1112).

Images crafted from brain scans make their way from the neuroimaging laboratory to journals, magazines, newspapers, the television screen, and other spheres of public circulation, where they make claims on the personhood of viewers (Dumit 2004). Through these images, people are encouraged to see themselves from a neuroscientific perspective where personhood is reducible to a normalized image of the brain's structure and biochemistry. Standardized brain images portray kinds of brains that define the categories to which individuals belong. Brain images labeled "normal controls," "schizophrenia," and "depression" seemingly illustrate a clear, persuasive, and visually unambiguous difference between normal and unhealthy brains. However, these images obscure the fact that individuals tested as controls because they show no problems on neuropsychological tests can nonetheless exhibit neuropathology in the brain scanner, and conversely that brain scans of individuals diagnosed with a psychiatric condition can surprisingly reveal normal looking brains (Ibid.).

Despite the variability, individuals who submit themselves to experimentation agree to the categories of the researcher and "act as homogenous subjects stereotypically responding to stimuli" (Roepstorff 2004, p. 1108). During analysis stages, variability across research subjects is frequently treated by scientists as

a source of noise and discarded by averaging data (Kanai and Rees 2011). Individual differences, however, can be a rich and important source of information that can be exploited to link brain anatomy to human cognition and behavior (Ibid.). Normalizing processes in scientific experimentation and research dissemination affect the way neuroscientists think about the brain and significantly challenge public understandings of neural, cognitive, and behavioral diversity. Inviting anthropologists to observe, document, and analyze these research dynamics might be a preliminary step in constructing a field of neuroscience that is aware of its pivotal role in shaping personhood and social values outside the laboratory. Developing ways to represent the heterogeneity of research subjects is vital to promoting social inclusion and discontinuing practices that essentialize, stereotype, and stigmatize.

To return to puzzles of our galaxy, it was Johannes Kepler (1571–1630) who first calculated the true motion of the planets. Diverted from his work, he spent the final years of his mother's life defending her against accusations of being a witch. During that period, witch-hunts were a means for the Church to sustain a singular hegemonic worldview. Those accused of witchcraft, like Johannes Kepler's mother Katharina Kepler, were usually guilty of little more than doing things a little differently. When science finally annexed religion to the rights of classifying the natural world, those in power had to navigate new pathways to sustain their hegemony. Natural philosophers started to be called "scientists," a term coined by William Whewell in the 1830s (Sardar 2000), and around 1840, a carpentry term for upright and perpendicular, the "norm," became the root for a constellation of words to refer to the common type or standard (Davis 1995). Terminology may have changed but have hegemonic injustices remained?

What Is Degenerate?

Etymologically, degeneration is simply an alteration in the structure of a preexisting form. From the Latin, *degeneratus*, something is said to degenerate when it "moves away from its genus or type, so that it is no longer general or typical" (Schwartzman 1994, p. 68). Prior to the nineteenth century, the word "degeneration" was used with both positive and negative connotations. The French naturalist Philibert Commerçon (1769), for example, saw degeneration as flowing from natural to civilized, while others used the word "degeneracy" to express their fears of civilization collapsing into chaos (Willard 1673; Sherwill 1704; Warne 1739). With the anxieties of productivist Christian colonial empires, the association of degeneracy with moral decline and negative dilapidation prevailed (Mason 2010).

The second half of the eighteenth century saw the writing of the Virginia Declaration of Rights, the Declaration of Independence, and *Les Droits de L'Homme et du Citoyen*. In a world that was beginning to question established hierarchies, increasingly debating human rights, and in some circles was fighting to hold on to an international slave trade, the political elite turned to the work of a rising class of natural philosophers and physicians as a means to maintain power.

Johann Friedrich Blumenbach, often cited as one of the founders of anthropology, compared different races and argued that people with black, yellow, brown, and red skin color were degenerates from the original white color (Blumenbach 1775). Such theories were readily seized on to legitimate racial inequalities, imperial conquests, and colonial exploitation.

The classification of "degenerate" was used to sideline ethnic groups as well as individuals with abnormal behavior. In the official statistics of the nineteenth century, statistics of deviance began to proliferate around 1820 (Hacking 1986, p. 222). Deviations from the mean "constituted ugliness in body as well as vice in morals and a state of sickness with regard to the constitution" (Vertinsky 2002, p. 101). Degeneracy became more than a word to describe an alteration in characteristics, it gained moral implications and was used to portray defect or decline (Hacking 2001, p. 144). Degenerates were people who were mad, criminals, prostitutes, vagrants, and those who committed suicide (Ibid.). The problem of seeing and classifying degenerates involved reasoning what should be done about them (see Rajchman 1988, p. 102). The growth of statistics of deviance coincided with the abundance of laws about crime and suicide (Hacking 2001, p. 143).

The early brain sciences claimed jurisdiction over the label "degenerate" once a biological etiology of degeneracy was constructed. It was Benedict Augustus Morel (1857) who placed psychiatry within the framework of medicine by arguing that madness was linked to biological alteration. In his seven hundred page volume entitled *Traité des dégénérescences physiques, intellectuelles et morales de l'espèce humaine*, Morel defined degenerations as "deviations from the normal human type which are transmissible by heredity." When the leadership of French psychiatry shifted over to Germany, Morel's ideas traveled quickly to the new center of psychiatric thought. Max Nordau, a German physician, popularized Morel's conceptualization of degeneracy in his widely translated book, *Entartung*, published in English as *Degeneration* (1968[1892]). Eugene Talbot followed suit in his 1898 book, *Degeneracy: Its Causes, Signs and Results*, where degeneracy was vociferously associated with contagious and infectious diseases, destructive behavior, toxic agents, unfavorable climate, mental decline, consanguineous and neurotic intermarriages, juvenile obesity, impure food, arrested development, skeletal anomalies, sensory deterioration, paranoia, hysteria, idiocy, and one-sided genius, as well as social parasitism, moral degradation, and cultural demise (Patrick 1899). An inflected theory of degeneracy that provided biological explanations for crime and mental illness was eventually swept up in eugenic theories. After World War I, degeneration theory became politically vicious and eugenic theory was implemented in the segregation and sterilization of degenerates (Lawrence 2009, 2010).

Degeneracy also became associated with drug use in 1857 (Aurin 2000). Drugs such as opium and marijuana were seen as a cause of degeneracy manifested as delinquency, and a lack of moral character. However, the classification of these drugs was more closely aligned with nineteenth century colonial elitism than with any deleterious physiological effects. Drugs used by the bourgeoisie (e.g., nicotine and caffeine) were seen as good, while drugs used by migrants and the lower classes

(e.g., opium and marijuana) were seen as bad. Subsequent research has shown that the use of illicit drugs is not causally associated with violence (Parker and Auerhahn 1998), the number of deaths in America from alcohol, tobacco, or poor diet and physical inactivity is higher than deaths from illicit drug use (Mokdad et al. 2004), and little evidence supports a link between illicit drug use and psychosis except among copiously habitual adolescent cannabis users who may double their risk of unmasking a genetic predisposition (Arsenault et al. 2004; Hall 2006; Macleod et al. 2004; Wada 2011).

The enactment of drug laws in the twentieth century was strongly linked with the formation of national identity (Weimer 2003) and catalyzed by racial anxiety derived from patterns of migration and social change. Anti-opium laws in America and Australia, for example, were a means to suppress Chinese immigrants during a period of economic recession, social tension, and class conflict (Hoffman 1990; Manderson 1999). Marijuana was criminalized in America not so much because of the effect of the drug but because of its consumption by a marginalized ethnic group, Mexican laborers, who were stereotyped as poor, violent, and criminal (Himmelstein 1983; Nadelman 1989). By using the neurophysiological effects of these drugs as a pretext, authorities were able to further marginalize ethnic minorities. As expected, mass media played a major role in linking illicit drugs with prostitution, poverty, and crime (Musto 1973).

While drugs of the lower classes were criminalized, substances of the elite escaped negative classification. Sugar is a prime example of a substance that through its elitist consumption and colonial trade resisted classification as a drug despite its addictive, health-harming, and drug-like properties (Lustig et al. 2012). Nicotine and caffeine are other examples of substances that for a long time escaped negative labels. Meanwhile, the criminalization of opium, marijuana, and other drugs benefited pharmaceutical companies that no longer needed to compete with unregulated self-medication (Musto 1973). Keeping bioactive substances available only on medical prescription covertly enabled pharmaceutical companies to exploit doctors as their unpaid sales force, while at the same time maintaining high prices for off-patent products that could be manufactured at minimal costs. The authoritative power of medical professionals became two-fold: They defined drug-users as degenerate and only condoned drug use under the strict surveillance of initiated professionals. Too often these classifications were inflected by social biases. In the same way that witch-hunts gave the Catholic Church an opportunity to flex and embolden its power, criminalizing drugs has been used to sustain a political, economic, and social hegemony.

The history of the conceptualization of degeneracy shows how badly diversity and variation has been understood. Using the word "degeneration" often implies that there is a more perfect state, and the term "degenerate" still conjures images of racism reminiscent of a time when difference was despised and misapprehended. The history of racism is a sensitive topic for anthropologists, because their field both founded scientific racism and pioneered the antiracist movement (Hill 1998). In the nineteenth century, the evolution of human societies was conceived as a hierarchical progression from primitive to technologically complex

(e.g., Tylor 1871; Morgan 1877). The blending of this form of evolutionism with the then-accepted ranking of racial groups (see Gould 1981) together with a theory of degeneracy that associated behavioral phenotypes with biogenetic factors (see Pick 1989) fuelled the growth of movements such as eugenics. When eugenicists and racial theorists adopted the concept of degeneracy, phenotypic variations became ranked according to cultural prejudice. Countering these movements were advocate anthropologists such as Franz Boas and his students who found the need to separate the idea of race and culture, and "argue against the unilineal progression of evolutionary stages" (González 2004, p. 19). To counter racism, racial classifications were distanced from biogenetic etiology by arguing that human behavior was conditioned by social and historical circumstances. Evolutionism was pushed to the periphery of some streams of anthropology and with it the concept of degeneracy became inadvertently ignored. Culture became a way to describe human diversity without recourse to biogenetic differences.

Many of the terms of common language, as D'Andrade (1995) observes, blend something about the world with our reaction to it. The term "degenerate" is no exception to D'Andrade's observation. The prevalent historical usage of the word "degenerate" in racial discourse, health disorders, and mental disease highlights normative assumptions of the common "type" and simultaneously demonstrates how difference and diversity have been misrepresented. Hacking (2001, p. 146) distinguishes between degeneracy associated with inherited deviant traits where "alcoholics beget alcoholics, male criminals beget male criminals, child abusers beget child abusers," and allotropic degeneracy meaning "any two or more forms in which a chemical element may exist; carbon, for example, may exist as coal, diamond, or the bucky balls named after Buckminster Fuller." For some nineteenth century behavioral theorists, degeneracy was not strictly uniform and could appear in any form across generations – as alcoholism in one generation, for instance, epilepsy in the next, and crime in the third. While Hacking describes allotropic degeneracy as the inheritance of a disorder with differing symptoms in each generation, allotropic degeneracy in the physical sciences is a neutral description. The many different lattice configurations that an element can form are simply alternative arrangements that exhibit divergent properties.

The attachment of the word "degeneration" to a legacy of moral issues obfuscated the important development of a value-free operational conceptualization of "degeneracy" by George Gamow, a Ukrainian-born American scientist, mathematician, and theoretical physicist (Mason 2010). Gamow contributed to the final solution of the coding problem of DNA by suggesting that the genetic code was degenerate. By degenerate, Gamow meant that some amino acids were recognizable by two or more nucleotide triplets. Degeneracy, as Gamow used the term based on his background in physics, referred to heteromorphic isofunctional elements. In addition to being a recognized characteristic of genetic codes (Goodman and Rich 1962; Reichmann et al. 1962; Weisblum et al. 1962; Barnett and Jacobson 1964; Mitchell 1968; Konopka 1985; McClellan 2000; Frank 2003; Gu et al. 2003), degeneracy thus understood is also known to be a feature of immune systems (Cohen et al. 2004; Fernandez-Leon et al. 2011), respiratory network regulation

of blood-gas homeostasis (Mellen 2010), human movement (Mayer-Kress et al. 2006; Barris et al. 2012), and brain structure (Tononi et al. 1999; Edelman and Gally 2001; Price and Friston 2002; Friston and Price 2003; Noppeney et al. 2004; Figdor 2010). In the brain, for example, different populations of neurons in response to identical external stimuli can produce similar behavioral responses (Noppeney et al. 2004). Gamow's obscure terminology, for many years, demoted the importance of his conceptual contribution to merely the methods sections of science papers. In more recent years, the applicability of the concept of degeneracy has become increasingly acknowledged with neural degeneracy even recognized as foundational to the ability of the brain to acquire cultural skills (Downey 2012).

Degeneracy is an important feature of complex and selectional systems. Without degeneracy, selective processes would not be possible, because, as Edelman and Gally (2001) have established, a population of structurally dissimilar traits is a necessary prerequisite for, an essential accompaniment to, and an inevitable product of the processes of selection. In the brain, degeneracy means that there is a many-to-one mapping between brain areas and behavioral function. "[M]ore than one combination of neuronal groups can lead a particular output, and a given single group can participate in more than one kind of signalling function" (Edelman 1989, p. 50). The example of a neurological lesion that appears to have little effect upon behavior within familiar contexts reveals the presence of degeneracy in the brain (Tononi et al. 1999). Unfortunately, the association of brain areas with the names of famous brain scientists – Broca's area, Wernicke's area, and Brodmann's areas – creates the illusion that each brain area has a singular function and that these areas are fixed and stable. Even though scientists may acknowledge that this one-to-one mapping is not true, many experiments in genetics, the neurosciences, and other biological sciences still subscribe to this essentialist and simplistic model. Mental disorders, for example, are commonly essentialized because "variation is being 'filtered,' to some extent, by particular neurobiological processes" (Adriaens and De Block 2013, p. 115). Essentialism might be a "valuable epistemology," but it is an "ontological error" due to the ascription of natural essences that discord with the phenotypic and genotypic heterogeneity of brain function (Ibid.). The structural variation underlying functional plasticity is a distributed property of complex adaptive systems that has been hidden in plain sight (Edelman and Gally 2001), commonly overlooked because of a reductionist bias (Atamas 2005; Whitacre 2010), and ignored because the term, degeneracy, is misleading (Mason 2010).

Multiple pathways can lead to the same outcome, and locating degeneracy at multiple levels of complexity is a fruitful means of mapping dynamic systems. Taylor (2001), for example, considers acute depression in working-class women in London not simply as the consequence of a biological alteration but as the heterogeneous construction of intersecting variables that take place along the multistranded life course of each individual. In a society in which women are expected to be the primary caregivers for children, the different kinds of causes and their interlinkages include but are not limited to: the loss of, or prolonged separation from, the mother when the woman was a child; a severe, adverse event in the year prior to the onset of depression; the lack of a supportive partner; and

persistently difficult living conditions. These distributed intersecting processes can contribute to the heterogeneous construction of acute depression at a particular moment. Taylor also highlights that the origins of depression are founded upon a variety of different arrangements of biological influences, social structures, and cultural expectations over time. For treatment, Taylor promotes a multi-pronged approach where counselors, social workers, social policy makers, and we could include neuropsychiatrists in his list, can all view their engagement as interlinked, with no particular agent offering a complete solution on their own. In the construction of depression, there are alternative routes to the same end. Outcomes are not endpoints but snapshots of ongoing engagements between intersecting processes. Elements imbricated in one construction process are implicated in many others. From Taylor's description, the situation of acute depression in working-class women in London can be conceived of as a degenerate system in the Gamowian sense. The net flux of a variety of interacting elements in multifarious arrangements can give rise to the same psychological condition of depression. Rather than finding out merely what the brain of someone with depression looks like, a degeneracy-based approach recognizes that a significant amount of genetic and neural processes are open to social and experiential modification.

Due to its association with eugenics in the interwar years and the Third Reich's "Final solution," most mainstream biologists distanced themselves from degeneration theory (Lawrence 2010). Since World War II, the word "degeneracy" has largely disappeared from biological discourse because of the association with harmful ideas that scorned diversity. Ironically, regardless of the best-intended attempts to correct and erase the dangerously conformist behavior that arose from a flawed misappropriation of degeneracy, our productivist and consumerist lifestyle has still nonetheless managed to replace diversity with standardization. The negative controls of Nazi murder and mutilation have been substituted by the positive controls of materialism and consumerism as the drivers of homogeny. Riefenstahl's NAZI propaganda film, *Triumph des Willens* (Riefenstahl 1935), could be likened to modern day MTV filmclips, and the choreographed movement choirs, called *Gemeinschaftstanz*, of the Ministry for Enlightenment and Propaganda (Kew 2001, p. 78) could be likened to popular contemporary aerobics styles such as Zumba, Body combat, and Tai-Bo. Selectively breeding the ideal body is unethical, but marketing products that supposedly assist individuals attain the perfect body is demonstrably profitable. Though categorizations of normalcy and degeneracy may be dismissed in rhetoric and are often pilloried for their reactionary, racist, and eugenicist subtexts, the notion of the standard type maintains a strong, subliminal, and enduring influence in prescriptive modes of education, medicine, and popular culture. Old words may be shunned, but normative practices persist nonetheless. Political correctness might shape our vocabulary, but it does not necessarily stop discriminative practices or standardizing processes from being realized. Distinguishing the history of degeneration employed by medical scientists (Lawrence 2009, 2010) from degeneracy as defined in systems biology (Mason 2010) is the first step toward understanding the heterogeneous construction of human experience.

Brain lesion and functional neuroimaging studies are showing that a particular function can be sustained by degeneracy within one brain and in degeneracy over subjects, i.e., cognitive performance can be produced by multiple systems within one subject and by different systems in different subjects (Noppeney et al. 2004). Degeneracy distributed across a population of individuals suggests that degeneracy operates at the cultural level. Anthropologists recognize that the key to understanding human cultural diversity is the study of how cultural traits persist and change over time (Mulder et al. 2006). The operational framework of degeneracy is a flexible conceptual tool to map the transmission and transformation of cultural traits over time. Describing degeneracy at the neural level and the cultural level might offer anthropologists and neuroscientists a shared vocabulary. If anthropologists working at the cultural level and neuroscientists working at the neurobiological level can find a common language, then they will be able to contribute collaboratively to a holistic and ethical study of the diversity of human experience.

Conclusion and Future Directions

We live in an increasingly brain conscious society. It is not surprising to hear someone describe their mood in terms of serotonin levels or a rival's behavior in terms of a neuropsychological condition, or to ask a friend who is acting silly if they have forgotten to take their medication. Are we becoming neurocentric? And if we are, what are the consequences? Without understanding the cultural lens through which we prioritize mental processes in explanations of human behavior, we are likely to overlook major biases in our thinking. Beyond a mere biography of language terms, this chapter has charted social and political dimensions entangled in the way science categorizes and classifies the world. Combining the epistemology of neuroscience with the self-reflexive methods of anthropology is the first step toward what Domínguez D. et al. (2010) have termed "a humanistic science." If technological developments are not accompanied by advances in our conceptual tools, then we will still be stuck in the age of witch-hunts, which will distract us from the real work of studying human brain and behavior.

The brain sciences occupy a capitalist post-Cartesian world where connecting the mind and the body has become a commodity that can be sold. Substances and body practices concerned with brain function are an open target for commodification. Food companies, for example, are constantly trying to fashion products high in omega-3 and with a low glycemic index that trigger gustatory reward mechanisms and simultaneously enhance mental performance. Pharmaceutical companies are explicitly in the game, and so too is online gaming, public media, and the self-help industry. With increasing attention paid to intelligence, health, and personal achievement, what you consume is on the sights of suppliers. By encouraging the public to conceptualize brain function according to a scientific discourse of normalcy and degeneration (in the moralistic sense), the mental health industry can simultaneously market itself as one of the good guys while improving sales. With due acknowledgment to the practitioners who deliver empowering services to vulnerable populations,

professionals must also be concerned for those individuals who feel inadequate because of cultural values promoted by a mental health industry that creates more categories of abnormality than it knows how to heal. Neuroanthropological perspectives that situate experimentation and medical practice within social settings, cultural values, and individual experience can facilitate in building ethical scientific practices. Placing an emphasis on diversity and shifting degeneracy from a direct localization in brain function to a distributed localization at intersecting neural, cultural, and environmental levels will play a key role in shaping equitable practices.

Professionals in the brain sciences can help to dispel myths about normalcy and degeneration at all levels of society. For over 150 years, medical professionals have claimed jurisdiction over the identification of degeneration irrespective of their capacity to redress the condition effectively or the social consequences of the label. While scientific thinking in the West provides protective mechanisms against culture bound syndromes found elsewhere, such as koro (a syndrome found in places like Northeast India (Sachdev 1985), South Celebes (Van Brero 1897), and remote Guangdong, China (Wen-Shing et al. 1988), involving the belief that one's genitals are retracting) or latah (found among some people in Indonesia and Malaysia (Bakker et al. 2013), latah is a heightened startle response with tic-like behaviors), to what extent has science weakened our resistance to culture bound features of schizophrenia, depression, and dissociative disorders? A neuropsychological substrate is undoubtedly at play, but locating degeneration in the biology of individuals dismisses multiple levels of intersecting variables and the manifold pathways of potential intervention. Adopting a value-free operational conceptualization of degeneracy allows us to model variable factors at multiple levels of complexity (Mason 2010). Neuroscience models based on degeneracy, not normalcy, combined with a particularistic approach will provide a more holistic and more humanistic account of the diversity of cultural experience and the neural bases of that experience (Domínguez et al. 2010, p. 141; Domínguez 2007). Furthermore, taking a self-reflexive approach and questioning the most basic assumptions of our scientific culture might serve to de-stigmatize diverse mental conditions, open new avenues of treatment, and cultivate a better-integrated society.

Cross-References

▶ Cognition, Brain, and Religious Experience: A Critical Analysis
▶ Determinism and Its Relevance to the Free-Will Question
▶ Developmental Neuroethics
▶ Dissociative Identity Disorder and Narrative
▶ Ethics in Psychiatry
▶ Explanation and Levels in Cognitive Neuroscience
▶ History of Neuroscience and Neuroethics: Introduction
▶ Human Brain Research and Ethics
▶ Impact of Brain Interventions on Personal Identity

▶ Mind, Brain, and Education: A Discussion of Practical, Conceptual, and Ethical Issues
▶ Neuroethics of Neurodiversity
▶ Neuroimaging and Criminal law
▶ The Contribution of Neurological Disorders to an Understanding of Religious Experiences
▶ Toward a Neuroanthropology of Ethics: Introduction
▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues
▶ What Is Addiction Neuroethics?

References

Adler, S. R. (2011). *Sleep paralysis: Night-mares, nocebos, and the mind-body connection*. London: Rutgers University Press.

Adriaens, P. R., & De Block, A. (2013). Why we essentialize mental disorders. *Journal of Medicine and Philosophy, 38*, 107–127.

Amundson, R. (2000). Against normal function. *Studies in History and Philosophy of Biological and Biomedical Sciences, 31*(1), 33–53.

Amundson, R., & Tresky, S. (2007). On a bioethical challenge to disability rights. *Journal of Medicine and Philosophy, 32*, 541–561.

Arsenault, L., Cannon, M., Wittton, J., & Murrah, R. M. (2004). Causal association between cannabis and psychosis: Examination of the evidence. *British Journal of Psychiatry*, *184*, 110–117.

Atamas, S. P. (2005). Les affinités electives. *Pour la Science, 46*, 39–43.

Aurin, M. (2000). Chasing the dragon: The cultural metamorphosis of opium in the United States, 1825–1935. *Medical Anthropology Quarterly, 14*(3), 414–441.

Bakker, M. J., van Dijk, J. G., Pramono, A., Sutarni, S., & Tijssen, M. A. (2013). Latah: An indonesian startle syndrome. *Movement Disorders, 28*(3), 370–379.

Barnett, W. E., & Jacobson, K. B. (1964). Evidence for degeneracy and ambiguity in interspecies aminoacyl-sRNA formation. *Proceedings of the National Academy of Sciences of the United States of America, 51*, 642–647.

Barris, S., Farrow, D., & Davids, K. (2012). Do the kinematics of a baulked take-off in springboard diving differ from those of a completed dive. *Journal of Sports Sciences, 31*(3), 305–313.

Becker, A. E., Burwell, R. A., Gilman, S. E., Herzog, D. B., & Hamburg, P. (2002). Eating behaviours and attitudes following prolonged exposure to television among ethnic Fijian adolescent girls. *British Journal of Psychiatry, 180*, 509–514.

Blacking, J. (1977). Towards an anthropology of the body. In J. Blacking (Ed.), *The anthropology of the body* (pp. 1–28). London: Academic Press.

Blumenbach, J. F. (1775 [1969]). *On the natural varieties of mankind*. New York: Bergman.

Cohen, I. R., Hershberg, U., & Solomon, C. (2004). Antigen-receptor degeneracy and immunological paradigms. *Molecular Immunology, 40*, 993–996.

Commerçon, P. (1769). [Letter to Joseph Lalande], Mercure de France, November.

Conrad, P. (1981). On the medicalization of deviance and social control. In D. Ingleby (Ed.), *Critical psychiatry: The politics of mental health* (pp. 102–119). Middlesex: Penguin.

D'Andrade, R. G. (1995). Moral models in anthropology. *Current Anthropology, 36*, 399–408.

Davis, L. J. (1995). *Constructing normalcy: The bell curve, the novel, and the invention of the disabled body in the nineteenth century*. London: Verso.

Domínguez, D. J. F. (2007). Neuroanthropology: The combined anthropological and neurobiological study of cultural activity. Ph.D. Thesis, The University of Melbourne.

Domínguez, D. J. F., Turner, R., Lewis, E. D., & Egan, G. (2010). Neuroanthropology: A humanistic science for the study of the culture-brain nexus. *Social Cognitive and Affective Neuroscience, 5*(2–3), 138–147.

Downey, G. (2012). Cultural variation in rugby skills: A preliminary neuroanthropological report. *Annals of Anthropological Practice, 36*(1), 26–44.

Dumit, J. (1997). A digital image of the category of the person: Pet scanning and objective self-fashioning. In G. L. Downey & J. Dumit (Eds.), *Cyborgs & citadels: Anthropological interventions in emerging sciences and technologies* (pp. 83–102). Santa Fe: School of American Research Press.

Dumit, J. (2004). *Picturing personhood: Brain scans and biomedical identity*. Princeton: Princeton University Press.

Edelman, G. M. (1989). *The remembered present: A biological theory of consciousness*. New York: Basic Books.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 13763–13768.

Epps, J. (1859). *Consumption: Its nature and treatment*. London: Sanderson.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology, 16*, 319–324.

Eysenck, H. J. (1953). *Uses and abuses of psychology*. Middlesex: Penguin.

Eysenck, H. J. (1975). *The future of psychiatry*. London: Methuen.

Fernandez-Leon, J. A., Acosta, G., & Mayosky, M. (2011). From network-to-antibody robustness in a bio-inspired immune system. *Bio Systems, 104*(2–3), 109–117.

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science, 77*(3), 419–456.

Foucault, M. (1965). *Madness and civilization*. New York: Pantheon.

Foucault, M. (1972). *Histoire de la Folie à l'âge classique*. Paris: Editions Gallimard.

Frank, S. A. (2003). Genetic variation of polygenic characters and the evolution of genetic degeneracy. *Journal of Evolutionary Biology, 16*, 138–142.

Friston, K., & Price, C. J. (2003). Degeneracy and redundancy in cognitive anatomy. *Trends in Cognitive Science, 7*(4), 151–152.

Galton, F. (1907). *Probability, the foundation of eugenics*. Oxford: Oxford University Press.

Goffman, E. (1961). *Asylums: Essays on the social situation of mental patients and other inmates*. New York: Anchor Books.

González, N. (2004). Disciplining the discipline: Anthropology and the pursuit of quality education. *Educational Researcher, 33*, 17–25.

Goodman, H. M., & Rich, A. (1962). Formation of a DNA-soluble RNA hybrid and its relation to the origin, evolution, and degeneracy of soluble RNA. *Proceedings of the National Academy of Sciences of the United States of America, 48*, 2101–2109.

Gould, S. J. (1981). *The mismeasure of man*. Harmondsworth: Penguin Books.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., & Li, W. H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature, 421*, 63–66.

Hacking, I. (1986). Making up people. In T. C. Heller & C. Brooke-Rose (Eds.), *Reconstructing individualism: Autonomy, individuality and the self in Western Thought* (pp. 222–236). Stanford: Stanford University Press.

Hacking, I. (2001). Criminal behavior, de-generacy and looping. In D. Wasserman & R. Wachbroit (Eds.),. *Genetics and criminal behavior* (pp. 141–168). Cambridge: Cambridge University Press.

Hall, W. (2006). The mental health risks of adolescent cannabis use. *PLoS Medicine, 3*(2), e39.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature, 466*(1), 29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–135.
Hill, J. H. (1998). Language, race, and white public space. *American Anthropologist, 100*, 680–689.
Himmelstein, J. L. (1983). From killer weed to drop out drug. *Contemporary Crises, 7*(1), 13–38.
Hoffman, J. (1990). The historical shift in the perception of opiates: From medicine to social menace. *Journal of Psychoactive Drugs, 22*, 53–62.
Illich, I. (1976). *Medical nemesis*. New York: Pantheon.
Kanai, R., & Geraint, R. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience, 12*(4), 231–242.
Kew, C. (2001). From weimar movement choir to nazi community dance: The rise and fall of rudolf Laban's Festkultur. *Dance Research: The Journal of the Society for Dance Research, 17*(2), 73–95.
Konopka, A. K. (1985). Theory of degenerate coding and informational parameters of protein coding genes. *Biochimie, 67*, 455–468.
Laing, R. D. (1960). *The divided self: A study of sanity and madness*. Chicago: Quadrangle.
Lawrence, C. (2009). Degeneration under the microscope at the fin se siècle. *Annals of Science, 66*(4), 455–471.
Lawrence, C. (2010). Historical keyword: Degeneration. *Lancet, 375*, 975.
Lustig, R. H., Schmidt, L. A., & Brindis, C. D. (2012). Public health: The toxic truth about sugar. *Nature, 482*, 27–29.
Macleod, J., Oakes, R., Copello, A., et al. (2004). Psychosocial and social sequelae of cannabis and other illicit drug use by young people: A systematic review of longitudinal, general population studies. *Lancet, 363*, 1579–1588.
Manderson, D. (1999). Symbolism and racism in drug history and policy. *Drug and Alcohol Review, 18*, 179–186.
Maugham, S. (1938). *The summing up*. Garden City: Doubleday, Doran & Company.
Mason, P. H. (2010). Degeneracy at multiple levels of complexity. *Biological Theory, 5*(3), 277–288.
Mayer-Kress, G., Yeou-Teh, L., & Newell, K. M. (2006). Complex systems and human movement. *Complexity, 12*(2), 40–41.
McClellan, D. A. (2000). The codon-degeneracy model of molecular evolution. *Journal of Molecular Evolution, 50*, 131–140.
Mellen, N. M. (2010). Degeneracy as a substrate for respiratory regulation. *Respiratory Physiological Neurobiology, 172*, 1–7.
Mitchell, W. M. (1968). A model for protein biosynthesis predicated on the concept of metastable states: A postulated role for genetic code degeneracy. *Proceedings of the National Academy of Sciences of the United States of America, 61*, 742–747.
Mokdad, H., Marks, J., Stroup, D., & Gerberding, J. (2004). Actual causes of death in the United States, 2000. *Journal of the American Medical Association, 291*, 1238–1245.
Morel, B. A. (1857). *Traité des Dégénérescences Physiques, Intellectuelles et Morales de l'Espèce Humaine et des Causes qui Produisent ces Variétés Maladives*. Paris: Bailliere.
Morgan, L. H. (1877). *Ancient society*. New York: Henry Holt.
Mulder, M. B., Nunn, C. L., & Towner, M. C. (2006). Cultural macroevolution and the transmission of traits. *Evolutionary Anthropology, 15*, 52–64.
Musto, F. D. (1973). *The American disease: Origins of narcotic control*. New Haven: Yale University Press.
Nadelmann, E. (1989). Drug prohibition in the United States: Costs, consequences, and alternatives. *Science, 245*, 939–947.
Noppeney, U., Friston, K. J., & Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy, 205*, 433–442.
Nordau, M. S. (1968 [1892]). *Entartung* [Degeneration]. Lincoln, NE: University of Nebraska Press.

Obeyesekere, G. (1985). Depression, buddhism and the work of culture in Sri Lanka. In A. Kleinman & B. Good (Eds.), *Culture and depression: Studies in the anthropology and cross-cultural psychiatry of affect and disorder* (pp. 134–152). Berkeley: University of California Press.

Ochs, E., & Izquierdo, C. (2009). Responsibility in childhood: Three developmental trajectories. *Journal of the Society for Psychological Anthropology, 37*(4), 391–413.

Oliver, M. (1983). *Social work with disabled people*. Basingstoke: Macmillan.

Parker, R. N., & Auerhahn, K. (1998). Alcohol, drugs, and violence. *Annual Review of Sociology, 24*, 291–311.

Patrick, G. T. W. (1899). Book review of degeneracy: Its causes, signs and results. By Eugene S. Talbot, M. D., D. D. S. The contemporary science series. London, Walter Scott/New York: Charles Scribner's Sons. 1898. *Science, 9*, 372–373.

Pick, D. (1989). *Faces of degeneration: A European disorder, c. 1848–c. 1918*. Cambridge: Cambridge University Press.

Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends in Cognitive Science, 6*, 416–421.

Quételet, A. (1835). *Sur l'homme et le développement de ses facultés: ou essai de physique sociale: Tome I & II*. Paris: Bachelier.

Quételet, A. (1973 [1842]). A treatise of man and the development of his faculties. (Reprinted in *Comparative statistics in the 19th century*). Germany: Gregg International.

Rajchman, J. (1988). Foucault's art of seeing. *October, 44*, 88–117.

Reichmann, W. J. (1964). *Use and abuse of statistics*. Middlesex: Penguin Books.

Reichmann, M., Markham, R., Symons, R., & Rees, M. (1962). Experimental evidence for degeneracy of nucleotide triplet code. *Nature, 195*, 999–1000.

Roepstorff, A. (1999). Det Neurale Menneske—et Antropologisk Perspektiv. In O. Hoiris, H. J. Madsen, T. Madsen, & J. Vellev (Eds.), *Menneskelivets mangfoldighed*. Moesgaard: Aarhus Universitetsforlag & Moesgaard Museum.

Roepstorff, A. (2002). Transforming subjects into objectivity — An "Ethnography of Knowledge" in a brain imaging laboratory. *FOLK, Journal of the Danish Ethnographic Society, 44*, 145–170.

Roepstorff, A. (2004). Mapping brain mappers, an ethnographic coda. In R. Frackowiak et al. (Eds.), *Human brain function* (pp. 1105–11017). London: Elsevier.

Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179*(4070), 250–258.

Riefenstahl, L. (1935) *Triumph des Willens*. Music by Herbert Windt. Leni Riefenstahl-Produktion and Reichspropagandaleitung der NSDAP.

Sachdev, P. S. (1985). Koro epidemic in North-East India. *Australian and New Zealand Journal of Psychiatry, 19*, 433–438.

Sanders, J. (1808). *Treatise on pulmonary consumption, in which a new view of the principles of its treatment is supported by original observations on every period of the disease*. London: Walker and Greig.

Sardar, Z. (2000). *Thomas Kuhn and the Science Wars*. Cambridge: Icon Books.

Schwartzman, S. (1994). *The words of mathematics: An etymological dictionary of mathematical terms used in english*. Washington, DC: The Mathematical Association of America.

Seligman, R., & Kirmayer, L. J. (2008). Dissociative experience and cultural neuroscience: Narrative, metaphor and mechanism. *Culture, Medicine and Psychiatry, 32*, 31–64.

Shakespeare, T. (2010). The social model of disability. In L. J. Davis (Ed.), *The disability studies reader* (pp. 264–271). New York: Routledge.

Sherwill, T. (1704). *The degeneracy of the present age as to principles (a sermon preach'd before the University of Cambridge, on Sunday June 25)*. Cambridge: Cambridge University Press.

Silverman, J. (1967). Shamans and acute schizophrenia. *American Anthropologist, 69*(1), 21–31.

Sobo, E. J. (1993). The sweetness of fat: Health, procreation, and sociability in rural Jamaica. In N. Sault (Ed.), *Many mirrors: Body image and social relations* (pp. 132–154). New Jersey: Rutgers University Press.

Szasz, T. S. (1962). *The myth of mental illness*. London: Secker and Warburg.

Szasz, T. S. (1963). *Law, liberty and psychiatry*. New York: Macmillan.

Taylor, P. (2001). Distributed agency within intersecting ecological, social and scientific processes. In S. Oyama, P. Griffiths, & R. Gray (Eds.), *Cycles of contingency: Developmental systems and evolution* (pp. 313–332). Cambridge, MA: MIT Press.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 96*, 3257–3262.

Treacher, A., & Baruch, G. (1981). Towards a Critical History of the Psychiatric Profession. In D. Ingleby (Ed.), *Critical psychiatry: The politics of mental health* (pp. 120–149). Middlesex: Penguin.

Tylor, E. B. (1871). *Primitive culture*. London: John Murray.

Van Brero, P. C. J. (1897). Koro, eine eigenthümliche Zwangsvorstellung. *Allgemeine Zeitschrift für Psychiatrie, 53*, 569–573.

Vertinsky, P. (2002). Embodying normalcy: Anthropometry and the long arm of William H.Sheldon's somatotyping project. *Journal of Sport History, 29*, 95–133.

Wada, K. (2011). The history and current state of drug abuse in Japan. *Annals of the New York Academy of Sciences, 1216*, 62–72.

Warne, J. (2003 [1739]). *The dreadful degeneracy of a great part of the clergy, the means to promote irreligion, atheism, and popery. To which is prefixíd, a letter to . . .Mr. George Whitefield*. (Reproduction of original from the British Library). Farmington Hills, MI: Thomson Gale.

Watters, E. (2010). *Crazy like us: The globalization of the American psyche*. New York: Free Press.

Weimer, D. (2003). Drugs-as-a-disease: Heroin, metaphors, and identity in nixon's drug war. *Janus Head, 6*(2), 260–281.

Weisblum, B., Benzer, S., & Holley, R. W. (1962). A physical basis for degeneracy in the amino acid code. *Proceedings of the National Academy of Sciences of the United States of America, 48*, 1449–1454.

Wen-Shing, T., Kan-Ming, M., Hsu, J., Li-Shuen, L., Li-Wah, O., et al. (1988). A sociocultural study of koro epidemics in Guangdong, China. *The American Journal of Psychiatry, 145*, 1538–1543.

Whitacre, J. (2010). Degeneracy: A link between evolvability, robustness complexity in biological systems. *Theoretical Biology and Medical Modelling, 7*, 6.

Willard, S. (1673). *Useful instructions for a professing people in times of great security and degeneracy (delivered in several sermons on solemn occasions)*. Cambridge, MA: MIT Press.

Section V

Neuroethics and Identity

Neuroethics and Identity

22

Françoise Baylis

Contents

Abstract

Identity is a concept of pivotal importance in neuroethics, especially in relation to recent and emerging advances in neuroscience that promise to effectively treat a wide range of motor and psychological disorders. A persistent worry is that modern neurotechnologies (including electroconvulsive therapy, psychosurgery, neural stimulation, psycho-pharmaceuticals, and stem cell transplantation) may negatively affect personal identity. This introduction briefly summarizes the papers in the section "Neuroethics and Identity," each of which addresses a particular facet of identity as the concept is, or might be, applied in neuroethics. In an effort to clarify the diverse understandings of the concept, the introduction concludes with summary statements on numerical identity, forensic identity, practical identity, narrative identity, and extended identity.

Introduction

In contemporary Western political and ethical theory, persons are generally thought of as autonomous, rational, self-aware, self-reliant, and self-interested beings.

F. Baylis
Faculty of Medicine, Novel Tech Ethics, Dalhousie University, Halifax, NS, Canada
e-mail: Francoise.baylis@dal.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_9,

Psychological or cognitive changes that alter these features of persons, be they a result of illness (such as depression, dementia, or obsessive compulsive disorder), or a result of brain interventions (such as psycho-pharmaceuticals, neurosurgery, or neurological stimulation), are thus morally salient. Indeed, changes that dramatically alter personality, behaviors, beliefs, and perceptions raise serious concerns about threats to personal identity, especially when those changes result from novel treatments or enhancements. There is not as yet, however, widespread agreement in the neuroethics literature on the meaning and scope of the concept of identity. While narrative approaches to identity appear to have significant appeal, the proponents of narrative identity offer different interpretations of the concept. The five chapters in this section on identity (four by philosophers and one by a legal scholar) elucidate some of these different perspectives.

Overview

The opening chapter by Catriona Mackenzie and Mary Walker carefully outlines at the outset, three distinct concepts of identity – numerical, practical, and narrative identity. As summarized by the authors, numerical identity usefully addresses questions of metaphysical essence, individuation, reidentification, and survival. Answers to these questions typically focus on theories of psychological continuity (with a particular focus on reason and self-awareness), or on theories of bodily or biological continuity (with a particular focus on biological life). With practical identity, the focus is on one's normative self-conception; identity is a concept that captures one's defining beliefs and values, motives, emotions, and so on. With narrative identity, the focus is on self-constitution over time.

This chapter has two discrete aims. The first aim is to clarify the concepts of identity, authenticity, and autonomy in order to better understand and evaluate the empirical data available from first-person reports of dramatic personality changes following neurotechnological interventions – changes that are experienced as either disorienting or liberating. The second aim is to show that a relational, narrative approach to identity and autonomy provides a more useful normative framework than an ethics of authenticity for assessing "the ethical implications of changes in identity arising from the use of neurotechnologies" (this volume p. 382). Mackenzie and Walker identify three problems with authenticity as the frame of reference for discussions of identity. First, the authenticity framework is conceptually misleading insofar as it fails to carefully demarcate authenticity as self-discovery from authenticity as self-creation. Second, each of these conceptions of authenticity is flawed insofar as authenticity as self-discovery wrongly presumes a static inner life and authenticity as self-creation blurs an important distinction between identity and autonomy. Third, both of these conceptions of authenticity fail to properly account for the dynamic, relational nature of identity. This chapter ends with a strong endorsement of a relational and narrative understanding of identity and autonomy.

The chapter by Marya Schechtman examines the identity question from a narrative perspective with reference to the illness of Dissociative Identity

Disorder, a condition where two or more distinct personalities/identities are alleged to share one body. In 1996, Schechtman published *The Constitution of Selves*. In this book, she criticizes traditional philosophical understandings of personal identity that focus on questions of reidentification and, having done so, she makes a positive contribution to the literature on identity in developing her influential Narrative Self-Constitution View of identity. According to Schechtman, "self-constitution must be part of a viable account of identity." (95) Accordingly, persons constitute themselves by developing an autobiographical narrative – a life story – "through which they experience and act on the world" (this volume p. 396). Their autobiographical narrative allows them to make sense of both past and anticipated events and experiences.

Initially, one might think that the Narrative Self-Constitution View of identity could allow for the possibility of more than one person in one body, as each alter could have something like a personal narrative (a distinct life story). Schechtman insists, however, that the autobiographical narrative of each alter would likely be somewhat deficient. She writes: "The fact that 'someone else' has controlled the body to which an alter 'returns' after a hiatus makes the standard kind of person-constituting narrative impossible" (this volume p. 399). As an alternative, Schechtman posits a plausible narrative of a single person with a rare disorder who is psychologically fragmented and in need of reintegration. She does not, however, suggest this as the definitive answer to the question of whether there can be more than one person in one body. Her aim in this chapter is simply to highlight the complexity of Dissociative Identity Disorder in relation to both issues of moral and criminal responsibility.

The next chapter, by Thorsten Galert, examines the impact of brain interventions on personhood and personal identity. The focus of this chapter, unlike the previous chapters, is on the identity (and other) effects of brain interventions, not the effects of illness on identity. For illustrative purposes, Galert focuses on case reports involving Deep Brain Stimulation for Parkinson's Disease. Deep Brain Stimulation is an invasive neurotechnology known to result in important and sometimes quite dramatic psychological changes.

At the outset, Galert distinguishes changes in personhood from changes in personal identity. On his view, a person's existence can properly be described as coming to an end if one of the prerequisites for being a person is lost as a consequence of a brain intervention (as when, for example, the effects of Deep Brain Stimulation undermine autonomy to such an extent that a person can no longer be held responsible for her actions). Most times, however, brain interventions that result in psychological changes merely result in personality changes, and only a subset of these personality changes are profound or significant enough to qualify as changes in personal identity.

Galert uses the distinctions between psychological changes, personality changes, personal identity, and personhood to make explicit the normative implications of the different ways in which the concept of personal identity is used in the neuroethics literature. If this concept is understood in numerical terms, then it follows that a change in identity means that a person "disappears." Understood in this way, a change in personal identity resembles a loss of personhood insofar as either consequence can only be considered a grave risk of brain interventions. On the other hand, if personal

identity is understood in narrative terms, then it is possible to describe changes that result from brain interventions in positive terms (perhaps especially when such interventions aim to change a person's identity for the better). Reasoning along these lines, Galert suggests that narrative accounts of identity (as posited by Schechtman and others) may be helpful in making sense of personality changes.

The fourth chapter by Rob Wilson and Bartlomiej Lenart represents a potentially dramatic shift in approaches to identity. The authors take as their starting point the important role that memory plays in psychological accounts of personal identity and in the extended mind thesis. On psychological accounts of personal identity, what matters is psychological continuity over time, hence the important role of memory. According to the extended mind thesis, memory can be distributed between the body and the world, a thesis supported by our increasing reliance on cognitive tools and structures outside the body (e.g., pen and paper, calculator, computer) in the activity of remembering.

Building on the extended mind thesis, Wilson and Lenart argue that narrative memory (a defining feature of psychological accounts of personal identity) need not be limited to individualistic recollection but can "incorporate the world beyond the individual" (this volume p. xxx). One example of this is cognitive offloading, where people rely on external environmental factors to lighten internal memory load. Other examples of how memory is physically extended beyond the body involve shared or co-constructed memories between two or more individuals, where the task of remembering is distributed. In the later instance, individual narratives constitutive of personal identity are entwined with collective remembering.

The relationship between individual narratives and collective remembering suggests to Wilson and Lenart that tracking a person's psychological identity over time may involve many minds, and that sometimes "identities are realized in the collective remembering of others" (this volume p. xxx). One implication of this extended account of identity is that even individuals with cognitive limitations that prevent them from tracking their own identities can nonetheless have extended identities in virtue of the remembering of others.

The final chapter is by Jennifer Chandler, a legal scholar with a particular interest in neuroscience. Her chapter examines issues at the intersection of neuroscience, personal identity, and law, with reference to two discrete theoretical approaches to personal identity – numerical identity and narrative identity.

In discussing numerical identity, Chandler notes that in assigning responsibility for criminal activity, the law typically relies on physical evidence (such as eye witness reports, fingerprints, and DNA) to identify the perpetrator(s), having assumed a one to one correspondence between the psychological and physical being. But what if the accused claims to have Dissociative Identity Disorder where there is more than one psychological being in one physical person? And, what if the host or dominant identity claims that an alter identity is responsible for the alleged crime? Is Dissociative Identity Disorder a single disordered identity (a problem with meta-memory functioning), or are there multiple distinct identities within one body (a fragmentation of memory)? How should the law respond?

Setting aside this example where numerical identity may be relevant in law, Chandler moves on to a discussion of narrative identity and draws the reader's attention to three distinct issues. The first issue concerns legally coerced consent to "self" changing brain interventions (interventions expected to alter cognitive functions and personality attributes). On occasion, the law pressures convicted criminals to accept treatment aimed at changing personality and behaviors (e.g., anti-androgen drug therapy for sex offenders) in exchange for a specific legal advantage (e.g., avoiding incarceration, or receiving a reduced sentence). The second issue concerns the disclosure requirements for informed consent to brain interventions that may alter the "self." A treatment intended to alter memory (e.g., propranolol for post-traumatic stress disorder (PTSD)) might also undermine a person's ability to testify in court about the event that caused the PTSD. Should this nonmedical risk of treatment be disclosed during the consent process? The third issue concerns legal responsibility for behaviors committed after "self" changing brain interventions. Imagine a patient who engages in criminal activity while under the influence of brain stimulation. Should the treating physician bear any legal responsibility? The chapter does not aim to authoritatively answer the questions posed, given jurisdictional variation in the applicable laws, but rather highlights the complexity of some of the issues at the intersection of neuroscience, personal identity, and law.

Conclusion and Future Direction

In closing, this section on personal identity explores various understandings of the concept, including numerical identity, forensic identity, practical identity, narrative identity, and extended identity. Numerical identity typically focuses on psychological continuity and/or biological continuity. Forensic identity underscores the importance of continuity of consciousness and highlights the relevance of having moral responsibility for one's actions. Practical identity depends upon one's normative self-conception in addressing questions about the self, about who one is. It captures those beliefs and values, motives, emotions, etc. that one perceives as defining one's nature. With narrative identity, the focus is on a self-constitution over time in creating one's life story, a story that must nevertheless meet certain external constraints, as not all stories qualify as identity-constituting narratives. And finally, with extended identity, the claim is that the psychological states constitutive of the identity of a person over time can extend beyond that person's body. An important, putative consequence of this understanding of identity is an expansion in the range of persons who can be recognized as having personal identity to include those with even extremely limited cognitive capacities.

Each of the contributors to this section on identity review one or more of these perspectives showing how each has something to say about the metaphysical and normative significance of personal identity when changes manifest as a result of illness or brain interventions. In different ways, all of the contributors embrace a dynamic understanding of identity and, as such, are less sympathetic to claims about threats to identity that are grounded in assumptions about authenticity.

Cross-References

▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Dissociative Identity Disorder and Narrative
▶ Ethical Implications of Cell and Gene Therapy
▶ Extended Mind and Identity
▶ Impact of Brain Interventions on Personal Identity
▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
▶ Neurotechnologies, Personal Identity, and the Ethics of Authenticity
▶ Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges

References

(2013). *Neuroethics*, 6(3).
Baylis, F. (2012). The self in situ: A relational account of personal identity. In J. Downie & J. Llewellyn (Eds.), *Relational theory and health law and policy* (pp. 109–131). Vancouver: UBC Press.
Baylis, F. (2013). "I am who I am": On the perceived threats to personal identity from deep brain stimulation. *Neuroethics, 6*, 513–526.
DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge: Cambridge University Press.
Elliott, C. (1999). *A philosophical disease: Bioethics, culture and identity*. New York: Routledge.
Lindemann Nelson, H. (2001). *Damaged identities, narrative repair*. New York: Cornell University Press.
MacKenzie, C., & Atkins, K. (Eds.). (2008). *Practical identity and narrative agency*. New York: Routledge.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Radden, J. (1996). *Divided minds and successive selves: Ethical issues in disorders of identity and personality*. Cambridge, MA: MIT Press.
Schechtman, M. (1996). *The constitution of selves*. Ithaca: Cornell University Press.
Shoemaker, D. (2008). *Personal identity and ethics: A brief introduction*. Ontario, Canada Broadview Press.

23 Neurotechnologies, Personal Identity, and the Ethics of Authenticity

Catriona Mackenzie and Mary Walker

Contents

Abstract

In the recent neuroethics literature, there has been vigorous debate concerning the ethical implications of the use of neurotechnologies that may alter a person's identity. Much of this debate has been framed around the concept of authenticity. The argument of this chapter is that the ethics of authenticity, as applied to neurotechnological treatment or enhancement, is conceptually misleading. The notion of authenticity is ambiguous between two distinct and conflicting conceptions: self-discovery and self-creation. The self-discovery conception of authenticity is based on a problematic conception of a static, real inner self. The notion of self-creation, although more plausible, blurs the distinction between identity and autonomy. Moreover, both conceptions are overly individualistic and fail sufficiently to account for the relational constitution of personal identity. The authors propose that a relational, narrative understanding of identity and

C. Mackenzie (✉) • M. Walker
Department of Philosophy, Macquarie University, Sydney, NSW, Australia
e-mail: catriona.mackenzie@mq.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_10,

autonomy can incorporate the more plausible aspects of both interpretations of authenticity, while providing a normatively more illuminating theoretical framework for approaching the question of whether and how neurotechnologies threaten identity.

Introduction

Developments in the cognitive neurosciences are rapidly expanding scientific knowledge of the structures and functions of the brain. They are also enabling the development of a range of neurotechnologies that directly manipulate the brain in order to improve patients' functioning or capacities, or alter mood, cognition, behavior, or personality. These neurotechnologies include electroconvulsive therapy (ECT), psychosurgery, neural stimulation, and psychopharmaceuticals. For example, neural stimulation, which includes transcranial magnetic stimulation and deep brain stimulation (DBS), can be used to treat a range of disorders, including motor disorders associated with Parkinson's disease (Schupbach et al. 2006), severe depression and, potentially, addictions, and other compulsive disorders (Pisapia et al. 2013). Psychopharmaceuticals, such as antidepressants, antipsychotics, and medication for attention-deficit hyperactivity disorder (ADHD), can also bring about global changes to functioning, mood, and behavior.

In the recent neuroethics literature, there has been vigorous debate concerning whether the use of such neurotechnologies is ethically problematic because, even if they improve the person's condition, they may change, or threaten, his/her identity. These concerns have been brought to the fore by empirical studies documenting such changes and patients' attitudes toward them. Carl Elliot, citing one such study, recounts the story of a neurologist for whom dinner party invitations triggered episodes of acute social anxiety (2003, pp. 55–56). After several weeks on an antidepressant, the patient describes himself as feeling "'like a new person'. He felt more confident, and began to enjoy initiating conversations with other people" (Elliot 2003, p. 56). On the other hand, in a study by Bolt and Schermer, a patient taking medication for ADHD reported that the treatment altered his personality in ways he disliked, making him cynical and indifferent (2009, p. 106).

Similar descriptions of changes to identity, character, and mood are reported in studies of patients treated with DBS for Parkinson's disease. A study by Schupbach and colleagues of 29 patients found that despite improvements in motor control, a number of patients experienced significant posttreatment difficulties with social adjustment (e.g., due to changes in relations with their spouses and families or in their professional lives) and self-perception, including with respect to body image (Schupbach et al. 2006). A female journalist who had hoped that DBS would enable her to get back to her life and renew her projects was reported to have lost interest in her work, her family, and her life after the treatment. Her loss of "vitality" appears to have been related to the loss of the goal that gave meaning to her life: fighting her disease. She is quoted as describing this change in the following

terms: “Now I feel like a machine, I’ve lost my passion. I don’t recognize myself any more” (Schupbach et al. 2006, p. 1812). Another patient, who described himself as having been “asleep” during the years of his illness, felt a renewed sense of confidence, energy, and interest in life as a result of the treatment. This change in the patient’s sense of self was, however, a source of marital conflict. His wife, who described them as “a perfect couple” when he was sick, seemed to interpret his current behavior as self-deceived: “‘Now, he wants to live the life of a young man, go out, meet new people,’” leading her to bemoan “‘I would rather he be like he was before, always nice and docile!’” (Schupbach et al. 2006, p. 1812).

Much of the ethical debate concerning the implications of these kinds of changes to identity has been framed around the concept of authenticity. This is no doubt due to Elliot’s influential critique of the increasing use of enhancement technologies, which he characterizes as quests for authenticity (2003). Elliot’s account of the ethics of authenticity is derived from Charles Taylor’s analysis of the sources of modern conceptions of identity (Taylor 1989; Elliot 2011). Just as Taylor understands his project as tracing the genealogy of the modern self, so the primary aim of Elliot’s work is cultural diagnosis: to disentangle “the complex relationship between self-fulfilment and authenticity, and the paradoxical way in which a person can see an enhancement technology as a way to achieve a more authentic self, even as the technology dramatically alters his or her identity” (2003, p. xxi). However, at times Elliot seems to move beyond diagnosis to endorse an ethics of authenticity, even if he is critical of enhancement technologies as means to achieve authenticity.

The language of authenticity has subsequently been taken up in the neuroethics literature and has become central to philosophical debates concerning the ethical implications of neurotechnologies, whether used for the purposes of treatment or enhancement. For example, David DeGrazia (2000, 2005) and Neil Levy (2011) appeal to the ideal of authenticity in defending the use of psychopharmaceutical enhancers, arguing that they can enable a person to be (or become) who he/she most wants to be. In the context of the use of DBS for treatment purposes, Felicitas Kraemer (2011b) notes that experiences of authenticity and alienation differ for different patients. Nevertheless, she suggests that “the concepts of authenticity and alienation are useful heuristic tools not only for a better philosophical understanding of the patients’ experience of DBS, but also, in some cases, the use of these concepts can even lead to a re-evaluation of the treatment and its side-effects and should, therefore, contribute to future reflection on DBS” (2011b).

Empirical studies of patients’ first-person phenomenological experiences of, and attitudes toward, identity changes arising from direct brain interventions are crucial for understanding some of the ethical issues raised by these neurotechnologies, whether used for purposes of treatment or enhancement. It is also true that patients’ first-person descriptions of these changes sometimes use language that seems to evoke the notion of authenticity. However, in our view the ethics of authenticity provides a misleading normative framework for assessing the ethical implications of changes in identity arising from the use of neurotechnologies. This is not merely because, as many in the literature have noted, there are significant differences

between the experiences of different patients with regard to whether neurotechnological interventions appear to enhance or hinder some sense of authenticity. Rather, it is because the ethics of authenticity, as applied to neurotechnological treatment or enhancement, is conceptually misleading. The notion of authenticity is ambiguous between two distinct and conflicting conceptions: self-discovery and self-creation (Levy 2007, 2011). The self-discovery conception of authenticity, as articulated in the neuroethics literature, is based on a problematic conception of a static, real inner self. The notion of self-creation, although more plausible, blurs the distinction between identity and autonomy. Moreover, both conceptions are overly individualistic and fail sufficiently to account for the relational constitution of personal identity.

The aims of this chapter are therefore twofold. First, in order to understand what might be at stake in some of the first-person patient reports of disorienting, or liberating, change consequent upon various neurotechnological interventions, it is important to clarify and distinguish the concepts of identity, authenticity, and autonomy. Second, we seek to show that the ethics of authenticity is misleading and individualistic. We also propose that a relational, narrative understanding of identity and autonomy can incorporate the more plausible aspects of the self-discovery and self-creation notions of authenticity, while providing a normatively more illuminating theoretical framework for approaching the question of whether and how neurotechnologies threaten "identity." It is important to clarify that the aim of the chapter is not to undertake first-order ethical analysis of specific neurotechnologies nor to weigh into debates about treatment versus enhancement; rather, it is to raise questions about the philosophical adequacy of framing these debates around an ethics of authenticity.

Section "Three Concepts of Personal Identity" distinguishes between three different concepts of personal identity (numerical, practical, and narrative), in order to clarify the sense of identity that is at stake in the concern that direct brain interventions might pose a threat to identity. Section "The Ethics of Authenticity: A Relational Critique" begins by explaining how this concern has been framed in debates about the ethics of authenticity, drawing on Levy's (2007, 2011) distinction between self-discovery and self-creation. We then explain why both conceptions of authenticity are premised on individualistic assumptions about identity, and why the notion of self-creation conflates identity and autonomy. In developing this critique, we propose a relational and narrative approach to identity and autonomy as an alternative to the ethics of authenticity.

Three Concepts of Personal Identity

When a person undergoes dramatic change of a kind that prompts the person or others to say "She is [or "I am"] not the same person any more," there are several ways of interpreting the kind of identity at stake. Drawing on the philosophical literature on personal identity, we can distinguish three concepts of identity that might be relevant to interpreting such statements: numerical, practical, and narrative.

Numerical Identity

Concepts of numerical identity seek to respond to questions concerning metaphysical essence, individuation, reidentification, and survival. Questions of metaphysical essence are concerned with the criteria for determining the essential features of the kind "person," that is, those features that distinguish persons from nonpersons. Questions of individuation are concerned with the criteria for determining what makes an individual person metaphysically distinct from other persons. Questions of reidentification are concerned with the criteria for determining what makes an individual numerically the same person despite qualitative changes over time or under different descriptions. Questions of survival are concerned with the criteria for determining when an individual person can be said to survive or to have ceased to exist. These questions are interconnected. How each question is answered will constrain possible answers to the other questions.

Philosophical debates concerning numerical identity are broadly divided into two kinds of answers to these questions. The first kind of answer, proposed by psychological continuity theorists, derives originally from John Locke (1975). Its most well-known contemporary exponent is Derek Parfit (1984). Psychological continuity theorists claim that what essentially characterizes the kind "person" is the psychological capacities characteristic of personhood, in particular capacities for reason and self-awareness. What individuates each person, as a metaphysically distinct entity, is their distinctive consciousness or mental contents; and what makes an individual the same person over time is continuity of this consciousness or mental contents. That is, what makes a person (A1) at one time the same as another person (A2) at a later time is that the right kind of psychological relations hold between them, in particular that A2 remembers doing or experiencing things done or experienced by A1, acts on intentions formed by A1; exhibits traits of character, personality, and temperament that are sufficiently similar to those of A1; and so on. A person survives on this view just so long as these relations of psychological continuity obtain.

Although psychological continuity theories seem intuitively plausible, much of this plausibility derives from our practical and evaluative interests in personal identity. As Marya Schechtman (1996) argues, these interests are connected to concerns related to moral responsibility (80–81), self-interested concern (83–85), survival (87–89), and compensation (86). We suggest that these are the kinds of interest that underlie concerns about whether direct brain interventions threaten identity, and as we explain below, we think that these concerns are better captured by practical and narrative approaches to identity. Moreover, as David DeGrazia argues, because numerical identity is concerned with the logical relation of identity, psychological theories face difficulties explaining the sense in which individual human persons are numerically identical to individual human biological organisms (2005, pp. 31–33, 45–46). The psychological continuity view implies, for example, that a person is not the same numerically distinct individual as the early stage fetus from which he/she developed, since no relations of psychological continuity hold between the person and the fetus. It also implies that an individual with dementia

who has very few remaining psychological connections with the person he/she once was, or a person in a persistent vegetative state (PVS) who has no psychological connections, does not survive as the same numerically distinct individual. Since the biological human organism persists in all these cases, the psychological continuity view thus seems to imply, problematically, that the person and the human biological organism are metaphysically distinct entities.

The second kind of answer, proposed by bodily or biological continuity theorists, rejects the claim that human persons are essentially persons (to use DeGrazia's terminology) and holds that, as far as metaphysical or numerical identity is concerned, we are essentially biological organisms, that is, human animals (Olson 1997; DeGrazia 2005). What individuates each human animal is their distinctive biological makeup, and what makes an individual the same over time is sameness of biological life. So long as this biological life survives, the human animal survives. On this view, personhood, rather than being essential to human persons, "represents merely a phase of our existence" (DeGrazia 2005, p. 49). So, the early stage fetus and the person he/she later becomes are numerically identical by virtue of the fact that they are the same continuing human animal. Likewise, the person who suffers dementia and eventually loses the capacities distinctive of personhood nevertheless survives as the same functioning biological organism or human animal.

Whichever kind of view provides the most plausible account of metaphysical or numerical identity, it is not numerical identity that is at stake when a person who has undergone some kind of direct brain intervention says that they no longer recognize themselves or that they feel alienated in some way from the person they have become – nor when a patient states that they finally (or once again) feel like "themselves." These people do not seem to be saying that they have become metaphysically distinct entities. Rather, what they seem to be saying is that aspects of their characters, interests, values, emotional responses, embodiment, or relationships with others have changed – sometimes in ways they do not like or identify with, sometimes in ways that they do. The concepts of practical and narrative identity provide accounts of the kind of identity that seems to be at stake in such statements.

Practical Identity

The concept of practical identity shifts the focus of concerns about identity from metaphysical to practical and evaluative questions. Rather than addressing questions concerning what a human person essentially is, or the necessary and sufficient conditions for reidentifying an individual as the same over time, theories of practical identity seek to address questions concerning "who" we are as persons, that is, what distinctively characterizes each of our individual first-person perspectives. Françoise Baylis provides the following list of such questions: "Who am I? Where am I from? Where have I been? Where am I going? What do I care about? What do I stand for? Who do I want to be? Who am I becoming?" (2012, p. 117).

Christine Korsgaard understands practical identity as a normative self-conception, "a description under which you value yourself, under which you find your life to be worth living and your activities to be worth undertaking" (1996, p. 101). In Korsgaard's view, our practical identities are both discovered and constructed. On the one hand, our normative self-conceptions are shaped by all sorts of factors which we did not choose and over which we have limited control. These include our gender, sexual, racial, cultural, linguistic, or ethnic identity; our bodily and intellectual capacities; our family relationships; and the social, historical, and political contexts in which we live our lives, all of which define who we are and what matters to us. On the other hand, insofar as we are agents, our identities are not simply given. Rather, we actively construct or constitute our identities through authorial processes of reflection, deliberation, and decision. By reflecting on and deliberating about who we are, what we value, and how we should act, we ask ourselves whether our identities, and the beliefs, values, motives, and emotions arising from them, should be normative for us, that is, whether they constitute sufficient *reasons* for action. If the answer is no, then we may seek to revise or change them. Practical identity is therefore both a condition for and a product of our agency.

The concept of practical identity provides a way of explaining a person's actions third-personally and also justifying them first-personally. For example, we can explain why a teacher dedicates so much of her time to class preparation by understanding that the role of being a teacher importantly informs her practical identity; or why a man makes considerable efforts to visit and keep in contact with his grandchildren who live interstate by understanding that being a grandfather is central to his practical identity. Third-personally, these identifications are explanatory. But first-personally, for the teacher or the grandfather, these reasons may be both explanatory and justificatory. That our reasons for action are expressive of our practical identities is why they are normative for us and so guide our actions.

The notion of practical identity can therefore explain why the wife of the man with Parkinson's disease who has dedicated herself to caring for her husband feels so disoriented by his renewed confidence, energy, and independence following DBS and wishes that he were "nice and docile" again – because she seems to have lost her value as a caregiver and so no longer seems to have a justification for her actions. It can also explain why the ADHD sufferer does not like the indifferent, cynical person he judges himself to have become as a result of treatment. These changes seem like threats to his practical identity because he does not value being the kind of person whose actions are guided by these attitudes.

Korsgaard's emphasis on the first-person perspective and the role of agency in constituting our practical identities is important. One problem with her account of practical identity, however, is that her analysis of self-constitution is primarily synchronic; that is, it focuses on moments of deliberation and decision but not on the constitution of our identities over time. Another problem is that her analysis of the connection between our practical identities and normative commitments is overly strong, suggesting that "to violate [these commitments] is to lose your integrity and so your identity, and to no longer be who you are" (1996, p. 102). Although some commitments might be strongly identity constituting in this sense,

these remarks suggest that Korsgaard conceives of our practical identities as somewhat rigid and static. This interpretation of the concept of practical identity is thus not well suited to understanding personal change over time and might seem to lend support to problematic notions of authenticity as being true to one's "real" inner self, as we discuss below in section "The Ethics of Authenticity: A Relational Critique."

Narrative Identity

In the recent neuroethics literature, the concept of narrative identity has become increasingly influential, including as a way of explaining the sense of identity that is at stake in the perceived threats to personal identity posed by direct brain interventions (see, e.g., Glannon 2008; Bublitz and Merkel 2009; Schermer 2009, 2011; Schechtman 2010; Baylis 2011; Johansson et al. 2011; Witt et al. 2011; Walker 2012). The concept of narrative identity, like the concept of practical identity, is responsive to first-personal questions concerning "who" we are as persons (Ricoeur 1992), or what Schechtman (1996) refers to as questions of characterization. However, whereas Korsgaard's notion of practical identity is primarily synchronic, narrative accounts of personal identity seek to explain the diachronic constitution and reconstitution of identity. They are thus better able to explain how we construct our identities in response to the flux, fragmentation, and contingency that are inherent in living a human life over time.

Narratives are holistic, organizing, interpretive structures. Good stories, for example, integrate the different events they recount into intelligible temporal sequences, enabling us to understand what happened, when, to whom, and why. A good literary narrative does more than this. It enables the reader to make sense of the actions and motives of the characters, their inner lives, and their relations to one another, by situating their discrete actions, motives, desires, and beliefs within the broader explanatory context of the narrative whole. According to narrative identity theorists, such as Schechtman (1996), we constitute our personal identities by developing self-narratives that function in similar ways as literary narratives. Self-narratives are implicit organizing and integrating structures that provide a lens through which, as persons, we interpret our personal histories, past actions, and experiences; project ourselves into our futures by shaping our intentions and plans; and understand our character traits, habits, and emotional dispositions. Because self-narratives are selective and interpretive, they enable us to make psychological and evaluative sense of our selves, forging patterns of coherence and psychological intelligibility in response to the changing and fragmentary nature of our lived experience.

The claim that persons constitute their identities through the construction of self-narratives may seem implausible for several reasons. First, unlike literary narratives, with carefully structured, teleological plots, human lives are subject to randomness and contingency and unfold and change in haphazard and unpredictable ways. The concept of a self-narrative therefore seems to suggest

that, as persons, we have much more authorial control over our lives and identities than we in fact do (see, e.g., Christman 2004 for a version of this criticism). Second, the notion of authorship seems to imply that rather than living our lives, we must be constantly self-consciously reflecting on them and, moreover, trying to fit them into the form of a recognizable literary genre (see, e.g., Strawson 2004 for a version of this and the following criticism). Third, a narrative involves interpretation rather than bare representation of our personal histories, raising the question of how narrative theories can distinguish between self-narratives that are truthful and those that are confabulated, self-deceptive, or paranoid (Kennett and Matthews 2012). Responding to these concerns will help to clarify further the notion of narrative identity and its relevance for neuroethics.

In response to the first and second worries, it is important to emphasize that the concept of narrative identity does not imply either that human lives must have the tight structure of a plot, or that rather than living we should be constantly self-reflecting. The point is rather that, in the words of Charles Taylor, persons are “self-interpreting animals” (Taylor 1985, p. 45); that is, self-interpretation is integral to the activity of living a human life. Further, our self-interpreting activities deploy capacities for narrative understanding – for identifying and constructing patterns of coherence and meaning. Although narrative self-interpretations need not take the form of tightly structured or articulate literary narratives or even stories, as Schechtman (1996, p. 114) argues, they should be articulable to some degree, for example, in response to others’ requests for explanation of a person’s actions, motives, or emotional responses. The concern that narrative identity is inconsistent with acknowledging the extent to which human lives are subject to randomness and contingency is, therefore, misplaced. Narrative understanding is a way of making sense of flux, contingency, and temporal change. Contingency – illness, accident, trauma, and bereavement – can derail even the most carefully planned life and challenge our self-narratives. To reconstruct our lives and reconstitute our identities in the face of such contingency, we need to find ways of incorporating it into our self-understandings and life stories. Thus narrative coherence is dynamic and provisional. The patterns of coherence within a person’s identity shift and change over time, in response to contingency, or changes to a person’s commitments and values, intimate relationships, or sociopolitical environment.

In response to the third worry, although self-narratives are dynamic and provisional, it is not the case that any old story a person tells about him or her self can count as a self-constituting narrative. Narrative identity theorists have proposed a number of different constraints on self-constituting narratives. Schechtman proposes that a self-constituting narrative must meet the “reality” constraint, which specifies that a person’s self-narrative must cohere with reality; it cannot be delusional, psychotic, based on gross factual errors, resistant to revision in light of contrary evidence, or inconsistent with others’ views about oneself (Schechtman 1996, pp. 119–128). Hilde Lindemann Nelson (2001) proposes that to be self-constituting, a self-narrative must be “credible,” meaning that it must have strong explanatory force, fit the evidence, and correlate with a person’s actions. On either of these views, confabulatory self-narratives are not identity constituting.

The notion of narrative identity can clarify, in a similar way to the notion of practical identity, the sense in which neurotechnologies that radically alter a person's mood, cognitive capacities, or behavior might be seen from a first- or third-person perspective as potential threats to identity. But because of its focus on self-constitution over time, and capacity to articulate the dynamic nature of identity, narrative identity provides a better theoretical framework for understanding the ethical issues raised by such identity changes. As Schechtman (2010) explains in discussing the use of DBS to treat Parkinson's disease, psychological changes that are brought about suddenly and in ways that seem to be beyond the authorial control of the patient can cause significant disruptions to their self-narrative, making it difficult to construct a coherent narrative that connects their present experiences and sense of self to their past: "In general, patients who report adjustment problems seem to have a hard time seeing the life they were leading after treatment as the continuation of the life they were leading before, and so find themselves having to reinvent themselves" (2010, p. 138). Because narrative identity is dynamic, however, the notion of narrative identity can also help to explain how such self-reinvention is possible. That is, if the patient, along with his or her close associates, can construct an integrated narrative that makes sense of radical psychological change within the broader context of the person's life and that connects this change with the person's plans, goals, and relationships – for example, the goal of battling illness: "A longer term narrative perspective will thus provide a viewpoint from which what might look like a narrative break, up close, can be seen as a small segment of a continuous and self-expressive life narrative" (2010, p. 138).

Some disruptions brought about by neurotechnological interventions are of course so profound that the kind of reinvention Schechtman discusses is not possible. Walter Glannon discusses the case of a patient for whom treatment by DBS, while it controlled the motor symptoms of Parkinson's disease, resulted in mania, leaving the patient with the devastating choice between being admitted to a nursing home because of serious physical disability despite intact cognitive and affective capacities or being admitted to a psychiatric ward because of his mania (2008, p. 290). In relation to this case, Glannon raises the following questions: "How much disruption can one's life narrative accommodate without threatening the integrity of the whole? Is there a threshold below which alteration of the psyche is substantial enough to alter the identity of the person?" (2008, p. 291). In this kind of case, narrative reinvention of the self may not be possible, but as Baylis suggests, this is not so much because the person's identity has been radically disrupted. Rather, it is because the patient "is no longer able to contribute meaningfully to the authoring of [his] life" because both the illness and the treatment have impaired his capacities for agency (2011).

The following section provides an argument in support of this claim by explicating the relationships between identity, agency, and autonomy. In so doing it also investigates the adequacy of an ethics of authenticity for assessing the ethical implications of changes in identity arising from the use of neurotechnologies.

The Ethics of Authenticity: A Relational Critique

The ethics of authenticity might seem to provide a compelling language for articulating the issues raised by the effects of DBS or psychopharmaceuticals on personal identity. One reason for this is that the ideal of authenticity as being "true to oneself" seems to resonate with both first- and third-person descriptions. As some of the empirical studies cited above show, patients often characterize these effects in terms of feeling either unlike themselves or alienated from their characteristics, or, alternatively, as feeling like themselves once again, or even as becoming who they really are for the first time. Another reason is that, as Elliot points out, the notion of authenticity as a normative ideal has a strong moral pull in our culture (2003, p. 34). The idea that one should strive to be oneself, to get in touch with one's "inner depths" or listen to one's "inner voice," seems to play a central role in our conception of what is involved in leading a good, meaningful, and fulfilling human life: "The ethic of authenticity tells us that meaning is not to be found by looking outside ourselves, but by looking inward. The meaningful life is an authentic life, and authenticity can be discovered only through an inner journey" (2003, p. 35; see also 2011, p. 369). As a corollary, we judge those who are inauthentic – who follow the crowd, or are overly conformist – as in some sense morally compromised or as leading lives that are less than satisfactory. The authentic life is thus culturally valued as a better life.

But what does "being true to oneself" mean? As Elliot's analysis makes clear, the ideal of authenticity involves many different, and sometimes conflicting, strands: not being a fake; looking inward rather than outward to discover one's true or core self; leading a fulfilled and meaningful life; reconciling the inner self with one's outward presentation; and also being flexible and engaging in activities of self-transformation. Insofar as Elliot's aim is cultural diagnosis and critique, his analysis of the somewhat protean shape of the ideal of authenticity is not problematic. The notion of authenticity becomes more problematic, however, when it is used (including by Elliot himself) to frame philosophical debates about neurotechnological treatment and enhancement. As Levy argues (2011), these debates have coalesced around two conflicting strands of the notion of authenticity: self-discovery, and self-creation or self-transformation (see also Bublitz and Merkel 2009 for a related categorization). Levy characterizes these as stemming from "rival outlooks on what it means to be human" (2011, p. 314) and argues that both outlooks are deeply culturally entrenched in our attitudes toward identity. The argument of this section of the chapter is twofold: that both conceptions of authenticity are problematic and premised on individualistic assumptions; and that a relational and narrative understanding of identity and autonomy can integrate the plausible aspects of each conception, while eschewing individualism.

The notion of authenticity as self-discovery admits of more and less plausible interpretations. In its less plausible variants, authenticity or being true to oneself is taken to mean discovering – and expressing in one's actions and behavior – one's essential nature, personality, or defining core characteristics. The self, on this view,

is understood as a static inner entity, a true or essential self, awaiting discovery. The process of discovery is variously interpreted as looking deep into oneself in solitary introspection, finding one's inner voice, and remaking one's life, character, or body to make them correctly express this "true self." In enhancement discourses, this can involve altering one's "mutable self-presentation to match the enduring inner self," stripping away and "remolding the outer body in conformity with the inner being" (Elliot 2003, pp. 20–21, 32). Elliot discusses as examples of this conception of self-discovery some of the discourses surrounding male to female sex reassignment surgery, body building, and breast reduction or augmentation surgery, which often represent body modification as a quest to become the true self one always was. Similarly, psychopharmaceuticals may be used in pursuing the attainment of a "true self" in psychological or emotional terms (Kraemer 2011a, p. 52; see also DeGrazia 2005, p. 208). Elliot himself does not seem to endorse the notion of an essential self awaiting discovery, but he does seem to endorse the idea that people have "true selves" or "relatively stable and coherent set(s) of mental and physical attributes" (2003, p. 50). Another theorist who seems to appeal to the notion of the true self in the context of debates about neurotechnological intervention is Kraemer (2011a, b). In her analysis of Schupbach and colleagues' (2006) study of DBS patients, Kraemer (2011b) argues that the ethical issues at stake in identity changes resulting from DBS can best be understood in terms of an ethics of authenticity, and seems to interpret authenticity in terms of a true, real, or ideal self.

In our view, the notion of a true, inner, essential self is implausible, not only because it fails to account for the dynamic, embodied, and narrative constitution of identity but also because it fails to account for the relational dimensions of identity – assuming that the true self is a deep, inherent self, intrinsic to the person, which would remain the same regardless of the person's embodiment, or their social, cultural, and historical situation. What discourses of self-discovery might, more plausibly, be attempting to capture, as Levy points out, is the recognition "that people *do* have dispositions and talents and personalities, which fit them better for some activities than for others, and which make some ways of life more fulfilling for them than others" (2011, p. 312). Such discourses might also be attempting to capture the extent to which our identities are shaped by biological, social, historical, and cultural factors over which, as agents, we exercise little or only limited control – such as our sexed embodiment; who our parents are; our linguistic, historical and cultural heritage; our susceptibility to genetically inherited diseases; and so on. We suggest below, however, that this more plausible reading of the notion of self-discovery is better understood in terms of a relational account of narrative identity.

The notion of authenticity as self-creation also admits of more and less plausible interpretations. The least plausible interpretation is a caricatured version of the existentialist notions of radical choice and radical freedom. DeGrazia (2000) and Levy (2011), for example, interpret Sartre as committed to the view that every choice we make involves a radical act of self-creation: "The authentic individual recognizes that literally *nothing* – not their genes, not their past history, not their social relationships or their talents and skills, not morality and not God – stands in the way of their self-creation" (Levy 2011, p. 311). The most plausible

interpretation, and the one that is most influential in neuroethics, is that of DeGrazia, who criticizes Elliot's conception of authenticity as based on a "misleading image of the self as 'given', static, something there to be discovered," arguing in contrast: "One can be true to oneself even as one deliberately transforms and to some extent creates oneself" (2000, p. 35).

DeGrazia's account of authenticity as self-creation draws on a narrative notion of identity and links self-creation to autonomy. He understands self-creation as "the conscious, deliberate shaping of one's own personality, character, other significant traits. . .or life direction" (2005, pp. 89–90). In his view, projects of self-creation arise from narrative identity or a person's sense of self. However, whereas a person need not take an active, self-directing role in shaping his or her self-narrative, the process of self-creation

> occurs when an individual takes an active role in authoring the biography, making it a *lived* biography. Rather than wondering about how later chapters will turn out, or wondering how they'll turn out but with no sense of controlling their direction, the self-creator endeavors to write those later chapters – and perhaps, in light of the evolving story, edit earlier chapters as different themes emerge as critical and self-discoveries put old details in a different light (2005, p. 107).

The conception of narrative identity implied by this quotation seems odd. On the one hand, DeGrazia characterizes narrative identity as a person's first-person interpretive "inner story" through which she comes to know and understand herself. He also regards having a self-narrative as a precondition for prudential planning and moral agency (2005, p. 89). On the other hand, in this quotation the agential dimensions of narrative identity are obscured, and the implication seems to be that a person's self-narrative could simply unfold without her actively authoring that narrative. Active authorship, DeGrazia seems to be suggesting, is characteristic of self-creation rather than narrative identity.

So what does self-creating authorship involve? DeGrazia acknowledges that self-creation does not require or imply "unlimited capacity for self-change and control over one's destiny" (2005, p. 91) and that projects of self-change are constrained by the inevitable dependencies and finitude of human life and shaped by a person's genetic makeup and embodiment, and by randomness and contingency. Thus "while self-creation is possible, the range of possibilities available to a person is both opened up and limited by other major factors and processes that shape our lives" (2005, p. 92). So how, according to DeGrazia, do we take an active role in authoring our biographies, thereby engaging in projects of self-creation? His answer is by making plans and setting goals about the kind of person we want to be or the kind of life we want to live and then acting to realize those plans and goals through our choices and actions. Not all projects of self-creation are authentic, however. A person may have set himself the goal of being a lawyer and may have worked hard for many years at university and in his first professional job in a corporate law firm, but then come to realize that the kind of life he is leading is not really "him" and that he does not identify with the goal that seems to have been motivating his choices and actions. He realizes that he has been somewhat self-deceived in the pursuit of this goal, seeking to satisfy other people's

expectations without adequately reflecting on whether it is really what he wants to do with his life. To be authentic, according to DeGrazia, a project of self-creation must be autonomous and honest (2005, p. 112): it must be expressive of the person's own preferences and values, and he must identify with it on the basis of careful reflection (2005, p. 102).

In contrast to Elliot, who worries about whether changes to a person's identity brought about by neurotechnologies compromise authenticity, DeGrazia argues that such changes count as authentic so long as they express the person's values and he or she reflectively identifies with these changes. He gives the example of Marina, a woman with a troubled family background and a history of failed relationships. Through hard work and perseverance, Marina has managed to achieve success in her career, but she is not satisfied with aspects of her character, in particular her tendency to be pensive and brooding. What she wants to be is a person who is "more outgoing, confident, and decisive professionally; less prone to feelings of being socially excluded, slighted, or unworthy of a good partner; and less obsessional generally" (2000, p. 35). Having briefly tried to effect these changes through psychotherapy, she asks her psychiatrist for a prescription for Prozac, which she thinks may bring about the desired result more cheaply and efficiently. DeGrazia argues that whether Marina uses psychotherapy or psychopharmaceuticals to effect the desired change is irrelevant in determining whether it is authentic or inauthentic. What matters is whether it is in line with her values and self-conception and whether she identifies with the person she is seeking to become posttreatment.

The concern of this chapter is not to address the question of whether identity changes brought about by enhancement, such as in the case of Marina, can be authentic. It is rather to assess the adequacy of the notion of authenticity for understanding the ethical implications of identity changes arising from neurotechnological interventions. In our view the self-creation interpretation of authenticity is more plausible than the notion of self-discovery, insofar as it is consistent with a dynamic and narrative account of identity, and emphasizes the agential, authorial dimensions of identity construction. However, the notion of self-creation is misleading for two reasons. First, it is premised on an overly individualistic conception of identity. Second, as DeGrazia acknowledges, the ethically relevant consideration is whether identity changes brought about by neurotechnological interventions are autonomous. The appeal to authenticity is redundant and blurs the distinction between narrative identity and autonomy.

DeGrazia appears to conceive of narrative identity primarily as a matter of self-ascription. In discussing the example of Marina, he says "it is ultimately up to Marina to determine what counts as Marina and what counts as not-Marina" and "if Marina is able to rid herself of traits with which she does not identify, and decides that the 'real Marina' does not have those traits, no one is in a position to correct her" (2000, pp. 37, 38). In a later discussion (2005, pp. 86–88), he acknowledges that others do in fact play a role in the constitution of narrative identity, in three main ways. First, some interpersonal relationships are central to our identities, and so those particular others play "starring roles" in one's self-narrative. Second, when others "mirror" or reflect back to us their images or conceptions of who we are, this

mirroring can consolidate (or distort) our self-narratives. Third, others play an important epistemic role in providing a reality check on our self-narratives, helping to identify self-narratives that are deluded or inaccurate. With respect to this third role, however, DeGrazia insists that, except in cases where our self-narratives do not meet Schechtman's reality constraint, our first-person self-narratives must always be regarded as more authoritative.

DeGrazia's account of the role of others in the construction of our self-narratives is, however, insufficiently relational and still adheres to a view of the self as an inner citadel, to which others may be admitted but only on one's own terms. It is true that others do play the three roles identified by DeGrazia. However, we endorse a more thoroughly relational view of identity, according to which our self-narratives are not discrete and self-contained inner stories. Rather, our self-narratives are constructed through interpersonal relationships and in the context of the larger social, historical, political, and cultural narratives within which we live our lives and seek to define and understand ourselves. As Mackenzie has argued elsewhere, "We are always already caught up in relations with others, even prior to birth, and we acquire identities and agency within a community of agents and are constrained by complex networks of social norms, institutions, practices, conventions, expectations and attitudes" (2008a).

This view integrates insights from feminist relational theory with the concept of narrative identity. According to feminist relational theory, self-identity is intersubjectively and socially constituted, in relations of dependence and interdependence, beginning with the infant's dependency on and earliest interactions with her caregivers. Annette Baier coins the term "second persons" to characterize the implications of this primary intersubjectivity; namely, that persons are "essentially successors, heirs to other persons who formed and cared for them" (Baier 1985, p. 85). This primary intersubjectivity, which is rooted in corporeal interactions with caregivers, is subsequently layered by more complex forms of intersubjectivity, which are made possible by cognitive and linguistic development, and by our participation in the social world. This complex intersubjective layering includes the way our identities are shaped by familial and personal relationships; by our embodiment; by social identity categories, such as those relating to gender, race, ethnicity, class, sexual orientation, and disability; and by the cultural, religious, political, and geographical communities into which we were born or to which we now belong. According to feminist relational theory, this intersubjective shaping of our identities can be enabling or damaging – although often it will be a mixture of both. It is enabling to the extent that it fosters a person's capacities to fashion an autonomous self-narrative; it is damaging if interpersonal relationships and social structures oppressively constrain the range of identity-constituting narratives a person can enact, or thwart the development and exercise of autonomy competence.

Feminist relational theorists hold that autonomy is a competence, involving a complex repertoire or suite of reflective skills, which may be developed and exercised to varying degrees and in different domains (Meyers 1989). Autonomy competence encompasses not just the minimal requirements of legal competence – understanding, minimal rationality, and the capacity to communicate one's

decision – but an array of complex competences. These include volitional skills, such as self-control and motivational decisiveness; emotional skills, such as the capacity to interpret and regulate one's own emotions; imaginative skills, required for understanding the implications of one's decisions and envisaging alternative possible courses of action; and capacities to reflect critically on social norms and values (Meyers 1989; Mackenzie 2000, 2002). Most importantly, relational theorists also hold that these autonomy skills emerge developmentally and are sustained and exercised in the context of significant social relationships. Hence, autonomy competence is a relationally or socially constituted capacity, which requires sustained interpersonal, social, and institutional scaffolding.

Because autonomy competence is relationally constituted, its development and exercise can be impaired by abusive, coercive, or disrespectful personal relationships and by oppressive social structures involving social relations of domination, exclusion, or marginalization (Mackenzie and Stoljar 2000; Friedman 2003; Oshana 2006). Social oppression can also impair autonomy by restricting the range of identity-constituting narratives available to members of oppressed social groups, either overtly or via oppressive stereotypes (Nelson 2001; Baylis 2012), and by refusals of uptake or social recognition (Anderson and Honneth 2005). Relational autonomy theorists claim that the internalization of oppressive social stereotypes, and social relations of misrecognition that deny members of oppressed social groups the *status* of being autonomous agents, can further impair autonomy by undermining a person's sense of him or her self as an autonomous agent. One way this can occur is by corroding self-evaluative attitudes of self-respect (regarding oneself as the moral equal of others), self-trust (the capacity to trust one's own convictions, emotional responses, and judgments), and self-esteem or self-worth (thinking of oneself, one's life, and one's undertakings as meaningful and worthwhile) (McLeod 2002; Anderson and Honneth 2005; Mackenzie 2008b).

The implication of feminist relational theory for narrative identity is that we cannot be self-creators but "are (and can only be) dynamic complex co-creations informed by the perspectives and creative intentions of others" (Baylis 2012, p. 118). This is because "one's first-personal perspective is always already an intersubjective, intercorporeal perspective situated in a material and cultural world" (Atkins 2008, p. 56). Constructing a narrative identity is therefore an ongoing negotiation between first-person self-ascriptions of identity, second-person recognition (or misrecognition) by others, and third-person identity ascriptions, such as ascriptions of personality, character, action attributions, and so on.

This relational negotiation of identity is evident in the case described by Schupbach and colleagues (2006) of the DBS patient, who claimed to feel more like himself following treatment but whose wife had a very different view of his identity change. This case raises the question, addressed by DeGrazia, of what weight and authority should be given to the person's self-ascriptions versus third-person ascriptions. In response to this case, Kraemer (2011b), appealing to the ethics of authenticity, seems to agree with DeGrazia that first-person ascriptions are always more authoritative. She characterizes this case as one in which the patient feels authentic because he has recovered his autonomy, which Kraemer interprets as

being "master of his own destiny," that is, being able to live his own life rather than "existing as someone else's heteronomous object of care" (2011b).

In contrast Baylis, who endorses a thoroughly relational conception of narrative identity, offers a more nuanced analysis of cases like this (although she does not discuss this specific case) (2011, 2012). Baylis proposes what she refers to as an equilibrium constraint on identity-constituting self-narratives: a balance between a person's first-person self-ascriptions and others' perceptions. On this view, a self-narrative is identity constituting if it achieves a "temporary (even if very fragile) stability" (2012, p. 123) between first-, second-, and third-person perspectives. Achieving this stability is an ongoing "interpersonal, communicative activity" (2012, p. 123). Baylis suggests that when people's identity claims are challenged by others, they have several options including the following: to try to project their self-narratives more successfully, to revise their projected self-narratives in response to others' reactions, to defer to others' perceptions, to reject others' interpretations of their character or actions, or to seek recognition of their identities within different communities of belonging. Self-interpretation thus involves "an iterative cycle of 'self'-perception, 'self-projection, 'other'-perception, and 'other'-reaction" (2012, p. 128).

In the case described by Schupbach and colleagues (2006), the patient seems both to reject his wife's interpretation of who he is and to want to seek out a different community of belonging in which he will be recognized for who he takes himself to be. This is what Baylis refers to as his "preferred self-narrative" (2011). But whether he will be able to enact or "perform" this self-narrative and achieve a temporary stability will not be entirely up to him, since it will be subject not only to uptake by others but also to the progress of the disease and the success of DBS in treating it. This is why achieving equilibrium requires finding a balance between how the patient perceives and understands himself, how others see and understand him, and the constraints of his lived embodiment (2011). In social contexts in which the beliefs and attitudes of others toward physical and psychological disability are stigmatizing and negative, achieving equilibrium will be a fraught and difficult process (with or without neurotechnological intervention), often resulting in feelings of self-alienation. Baylis argues, however, that this sense of self-alienation should not be understood as the result of a threat to identity caused by disability or neurotechnological intervention. Baylis does not deny that life-changing disruptions to a person's narrative identity "undeniably constrain[s] the dialectical process of identity formation (and thereby alters a planned or anticipated narrative)" (2011). However, in her view, such experiences of self-alienation are usually either the result of the internalization of social stereotypes and social relations of misrecognition or the effect of direct brain interventions that impair a person's capacities for autonomous agency (2011).

Baylis' equilibrium constraint and her analysis of the likely reasons for the experiences of self-alienation that are sometimes described in patient narratives seem plausible in our view. Combining Baylis' analysis with the relational analysis of autonomy sketched out above, it should now be clear why the ethics of authenticity blurs the distinction between narrative identity and autonomy. The salient

issue, as DeGrazia's notion of authenticity as self-creation implicitly acknowledges, is not whether such interventions threaten identity, but whether they impair autonomy competence. However, by framing debates about neurotechnological interventions in terms of potential threats to or enhancements of identity, the ethics of authenticity obscures this question. Once this question is brought to the fore, however, it is evident that it can only be answered case by case. In some cases, such as in the tragic case discussed by Glannon and mentioned at the end of the preceding section, an intervention such as DBS can disrupt a person's autonomy competence to such an extent that he is unable to engage in narrative self-revision. In other cases, neurotechnological interventions, by alleviating the physical or psychological effects of illness (including mental illness), may thereby restore some of the volitional, emotional, motivational, imaginative, and critically reflective capacities necessary for autonomous deliberation and action. In so doing, such interventions may make it possible for a person to reengage in the process of reconstructing or repairing an integrated narrative identity. Relational theory shows, however, that a person's ability to do so is not just up to the person him or herself; it is also a matter of the extent to which others enable him/her to do so, by providing the social scaffolding and recognition required for autonomous self-narration.

Conclusion

This chapter has shown that in order to understand what might be at stake in phenomenological descriptions of disorienting, or liberating, change consequent upon various neurotechnological interventions, it is important to clarify and distinguish the concepts of identity, authenticity, and autonomy. The first section distinguished three different concepts of personal identity – numerical, practical, and narrative – and argued that the concept of narrative identity enables us to make best sense of these first-person descriptions. The second section showed why the ethics of authenticity provides a misleading normative framework for understanding the ethical issues at stake in neurotechnological interventions. First, the notion of authenticity, as it has been understood in neuroethical debates, is ambiguous between self-discovery and self-creation. Second, the notion of self-discovery is implausible. The notion of self-creation is more plausible insofar as it draws on a dynamic, narrative conception of identity. However, it blurs the distinction between identity and autonomy and so obscures the central ethical issue it seeks to address. Third, both interpretations of authenticity fail to account for the relational dynamics of identity and autonomy. We proposed that a relational and narrative account of identity and autonomy can incorporate the most plausible aspects of the ethics of authenticity and can explain what is at stake in first-person phenomenological descriptions of self-alienation. Such descriptions, as Baylis argues, draw attention to the distress experienced as a result of being unable to achieve equilibrium in one's self-narrative. This distress, in our view, points to threats to autonomy rather than to identity or authenticity, as the salient concern underlying narratives of self-alienation.

Cross-References

- ▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
- ▶ Dissociative Identity Disorder and Narrative
- ▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
- ▶ Ethics of Pharmacological Mood Enhancement
- ▶ Extended Mind and Identity
- ▶ Feminist Ethics and Neuroethics
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
- ▶ Reflections on Neuroenhancement

References

Anderson, J., & Honneth, A. (2005). Autonomy, vulnerability, recognition and justice. In J. Christman & J. Anderson (Eds.), *Autonomy and the challenges to liberalism* (pp. 127–149). Cambridge: Cambridge University Press.

Atkins, K. (2008). *Narrative identity and moral identity*. New York: Routledge.

Baier, A. (1985). *Postures of the mind: Essays on mind and morals*. Minneapolis: University of Minnesota Press.

Baylis, F. (2011). "I am who I am": On the perceived threats to personal identity from deep brain stimulation. *Neuroethics*. doi:10.1007/s12152-011-9137-1.

Baylis, F. (2012). The self in situ: A relational account of personal identity. In J. Downie & J. Llewellyn (Eds.), *Being relational: Reflections on relational theory and health law* (pp. 109–131). Vancouver: UBC Press.

Bolt, I., & Schermer, M. (2009). Psychopharmaceutical enhancers: Enhancing identity? *Neuroethics, 2*, 103–111.

Bublitz, J. C., & Merkel, R. (2009). Autonomy and authenticity of enhanced personality traits. *Bioethics, 23*(6), 360–374.

Christman, J. (2004). Narrative unity as a condition of personhood. *Metaphilosophy, 35*(5), 695–713.

DeGrazia, D. (2000). Prozac, enhancement, and self-creation. *Hastings Center Report, 30*(2), 34–40.

DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge: Cambridge University Press.

Elliot, C. (2003). *Better than well: American medicine meets the American dream*. New York: Norton.

Elliot, C. (2011). Enhancement technologies and the modern self. *Journal of Medicine and Philosophy, 36*, 364–374.

Friedman, M. (2003). *Gender, autonomy, politics*. New York: Oxford University Press.

Glannon, W. (2008). Stimulating brains, altering minds. *Journal of Medical Ethics, 35*(5), 289–292.

Johansson, V., Garwicz, M., Kanje, M., Schouenborg, J., Tingstrom, A., & Gorman, U. (2011). Authenticity, depression, and deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*(21), 1–3.

Kennett, J., & Matthews, S. (2012). Truth, lies, and the narrative self. *American Philosophical Quarterly, 49*(4), 301–316.

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.

Kraemer, F. (2011a). Authenticity anyone? The enhancement of emotions via neuro-psychopharmacology. *Neuroethics, 4*, 51–64.

Kraemer, F. (2011b). Me, myself and my brain implant: Deep brain stimulation raises questions of personal authenticity and alienation. *Neuroethics*. doi:10.1007/s12152-011-9115-7.

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Levy, N. (2011). Enhancing authenticity. *Journal of Applied Philosophy, 28*(3), 308–318.

Locke, J. (1975). *An essay concerning human understanding*. Oxford: Oxford University Press.

Mackenzie, C. (2000). Imagining oneself otherwise. In C. Mackenzie & N. Stoljar (Eds.), *Relational autonomy: Feminist perspectives on autonomy, agency and the social self* (pp. 124–150). New York: Oxford University Press.

Mackenzie, C. (2002). Critical reflection, self-knowledge and the emotions. *Philosophical Explorations, 5*(3), 186–206.

Mackenzie, C. (2008a). Introduction: Practical identity and narrative agency. In K. Atkins & C. Mackenzie (Eds.), *Practical identity and narrative agency* (pp. 1–28). New York: Routledge.

Mackenzie, C. (2008b). Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy, 39*, 512–533.

Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency and the social self*. New York: Oxford University Press.

McLeod, C. (2002). *Self-trust and reproductive autonomy*. Cambridge, MA: MIT Press.

Meyers, D. (1989). *Self, society and personal choice*. New York: Columbia University Press.

Nelson, H. L. (2001). *Damaged identities, narrative repair*. Ithaca: Cornell University Press.

Olson, E. (1997). *The human animal: Personal identity without psychology*. Oxford: Oxford University Press.

Oshana, M. (2006). *Personal autonomy in society*. Aldershot: Ashgate.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Pisapia, J. M., Halpern, C. H., Muller, U. J., Vinai, P., Wolf, J. A., Whiting, D. M., Wadden, T. A., Baltuch, G. H., & Caplan, A. L. (2013). Ethical considerations in deep brain stimulation for the treatment of addiction and overeating associated with obesity. *AJOB Neuroscience, 4*(2), 35–46.

Ricoeur, P. (1992). Life in quest of narrative. In D. Wood (Ed.), *On Paul Ricoeur: Narrative and interpretation* (pp. 20–33). London/New York: Routledge.

Schechtman, M. (1996). *The constitution of selves*. Ithaca: Cornell University Press.

Schechtman, M. (2010). Philosophical reflections on narrative and deep brain stimulation. *Journal of Clinical Ethics, 21*(2), 133–139.

Schermer, M. (2009). Changes in the self: The need for conceptual research next to empirical research. *The American Journal of Bioethics, 9*(5), 45–47.

Schermer, M. (2011). Ethical issues in deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*, 1–5.

Schupbach, M., Gargiulo, M., Welter, M. L., Mallet, L., Behar, C., Houeto, J. L., Maltete, D., Mesnage, V., & Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*, 1811–1816.

Strawson, G. (2004). Against narrativity. *Ratio* (new series), XVII, 428–452.

Taylor, C. (1985). *Human agency and language: Philosophical papers* (Vol. 1). Cambridge: Cambridge University Press.

Taylor, C. (1989). *Sources of the self*. Cambridge, MA: Harvard University Press.

Walker, M. J. (2012). Neuroscience, self-understanding, and narrative truth. *AJOB Neuroscience, 3*(4), 63–74.

Witt, K., Kuhn, J., Timmerman, L., Zurowski, M., & Woopen, C. (2011). Deep brain stimulation and the search for identity. *Neuroethics*. doi:10.1007/s12152-011-9100-1.

Dissociative Identity Disorder and Narrative

24

Marya Schechtman

Contents

Abstract

Dissociative identity disorder (DID) has been a subject of fascination to clinicians, philosophers, and the general public for well over a century. There are many reasons for this, not the least of which is the way in which this disorder challenges ordinary understandings of personal identity. It is not easy to say how many people are present in an encounter with a DID patient. Since facts about personal identity are intimately connected to prudential reasoning and the assignment of moral responsibility, perplexity about identity in these cases has potentially important practical implications. This chapter offers a theoretical framework for the fruitful investigation of questions of personal identity in DID in the form of the Narrative Self-Constitution View which argues that individuals constitute themselves as persons by developing and operating with an implicit autobiographical narrative in which they apply the normative constraints of personhood to their own lives. The Narrative Self-Constitution

M. Schechtman
University of Illinois at Chicago, Chicago, IL, USA
e-mail: Marya@uic.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_113,

View does not deliver a definitive answer to the question of how many persons one is confronted with in encountering a DID patient, but it does provide important insights into why a definitive answer should not be expected and sheds light on the role of embodiment in personal identity. The further value of this approach is demonstrated by showing how it can illuminate questions of moral and criminal responsibility in DID.

Introduction

Dissociative identity disorder (DID), formerly multiple personality disorder (MPD), has been a subject of fascination to clinicians, philosophers, and the general public for well over a century. There are many reasons for this, not the least of which is the way in which this disorder challenges ordinary understandings of *personal identity*. It is not easy to say how many people are present in an encounter with a DID patient. Since facts about personal identity are intimately connected to prudential reasoning and the assignment of moral responsibility, perplexity about identity in these cases has potentially important practical implications.

This chapter focuses on issues of personal identity in DID, offering a theoretical framework for exploring these questions in the form of a narrative account of personhood. The first step is to get a slightly more developed picture of the phenomenon of DID and of some of the outstanding disputes about the nature of the disorder. Next a *narrative* account of personal identity – the Narrative Self-Constitution View– will be introduced and applied to the case of DID. The Narrative Self-Constitution View does not provide a straightforward answer to the question of how many persons are present in an encounter with a DID patient, but it does offer important insight into this complex disorder, showing why a straightforward answer is not to be expected. The usefulness of this insight will be demonstrated by looking at the way in which the Narrative Self-Constitution View can illuminate tricky questions about moral and criminal responsibility that arise in cases of DID.

Some Preliminaries

According to the DSM-IV DID is characterized, among other diagnostic criteria, by "two or more distinct identities or personality states," at least two of which "recurrently take control of the person's behavior." Patients also demonstrate an "inability to recall important personal information that is too extensive to be explained by ordinary forgetfulness," and experience frequent blackouts. Slight modifications of these criteria have been proposed for DSM-V, but these will not be important for present purposes.

A general picture of what DID is like is familiar from popular depictions such as those found in *Sybil* and *The Three Faces of Eve*. The differences between

personality states in DID can be quite profound. The distinct identities or "alters" may be vastly disparate in temperament, values, and commitments. Usually, there are amnesiac barriers running in at least one direction, so that one alter lacks memory of what another has said or done, resulting in the blackouts, forgetfulness, and lost time that characterize the disorder (Dorahy 2001). Frequently there is a "host" personality that represents the core presentation which is interrupted by alters. The alters may live quite different lives, engaging in different activities and interacting with different groups of people. They may claim to be distinct individuals and may describe themselves as being different ages or sexes from one another and from the human body through which they are expressed. There are also reports of alters differing from one another in various physiological characteristics. It is claimed, for instance, that alters may have optical differences (one needing corrective lenses where another does not), or different allergic responses, or handedness. Some research has found that alters also differ in EEG and other patterns of brain activity (Coons 1988; Reinders et al. 2006).

The question of just how deep the differences between alters run is a contested one, however, and there has been a fair bit of work aimed at showing that the divisions among distinct personalities in DID are not as sharp as they are often taken to be. Studies have shown that although there are amnesiac barriers between alters, there is more leakage than there is typically thought to be. In particular, if we concentrate on types of memory other than autobiographical (the learning of skills, for instance) or on priming effects, there seems to be a good deal of carryover between alters, suggesting that the psyches of the individual personalities are not as hermetically sealed as it might at first seem (Eich et al. 1997). There have also been charges that claims of physiological differences among alters are exaggerated and anecdotal (Merckelbach et al. 2002).

Another aspect of the disorder that is a matter of deep controversy is its etiology. *DID* is widely believed to be a posttraumatic developmental pathology. A standard theory is that dissociation occurs as a protective response to childhood abuse or trauma. Alters develop when a traumatized child repeatedly compartmentalizes overwhelming emotions and memories, ultimately resulting in distinct personalities designed to deal with these powerful feelings (e.g., Loewenstein and Putnam 1990). This account enjoyed a great deal of popular attention during the 1990s with several dramatic and well-publicized stories involving "recovered memories" of childhood sexual abuse. In these cases, flashback memories of abuse or other trauma resurfaced in adulthood. Such memories were often associated with a clinical presentation of DID and often occurred during hypnosis or in the course of other therapeutic interventions. There has, however, been a strong challenge posed to the idea of recovered memory in general and to the etiology of DID in childhood trauma in particular. The "memories" recovered, it is argued, are the result of hypnotic, therapeutic, or cultural suggestions on a vulnerable mind. While claims of remembered trauma may be quite sincere, opponents argue, they are nevertheless unreliable. This has led some detractors from the standard view to question the very phenomenon of DID as a discrete syndrome (Loftus 1993; Piper and Merskey 2004).

Questions about the actual distinctness of the alters have a clear potential to impact the assessment of how many persons are present in a DID patient, and although somewhat more remotely connected, questions of etiology may also enter into these judgments. Conclusions drawn while these issues remain in question must therefore be considered provisional, but it is possible nevertheless to make quite a bit of progress in understanding how judgments of personal identity can be made in the face of this perplexing disorder. The next section introduces an account of personhood and personal identity that will contribute to such progress.

The Narrative Approach

There are, of course, many different theoretical accounts of *personhood* and *personal identity*, most of which have direct implications for the question of how many persons are present in a case of DID. Views that define identity in *narrative* terms are especially illuminating in this context. There are a wide range of different views that fall under this general rubric, and they differ in important respects (Baylis 2012; Bruner 1990; Dennett 1992; Goldie 2012; Ricoeur 1992; Taylor 1989). What they all have in common is the claim that it is illuminating to think about persons and their identities in narrative terms. The version of this approach employed here will be the Narrative Self-Constitution View developed by Schechtman (1996, 2011). According to this view, persons constitute themselves by developing a largely implicit autobiographical narrative which serves as the lens through which they experience and act on the world. The idea is not a complicated one: basically as children mature, they learn to think about the events in their lives as connected to one another by a certain kind of logic involving both natural causes and human motivations. This way of thinking about oneself and one's life has both phenomenological and behavioral effects, and these effects constitute personhood.

Here is an example of the relevant phenomenon: take an event described neutrally as stepping off of an airplane at a particular airport. For the person leaving the plane, this event will be very different if she is (a) arriving at the country where she will finally pick up the child whose adoption she has eagerly anticipated, (b) arriving at the country of her deployment for dangerous military duty having left small children behind at home, (c) arriving at the place where she will give an academic address and receive a prestigious reward, or (d) arriving for the funeral of a beloved relative who died unexpectedly and very young. The same neutrally described event will yield a different experience and lead to different actions and decisions in each of these cases because it is part of a different narrative and it takes its character from the story of which it is a part.

The Narrative Self-Constitution View claims that all of our experiences are like this. As persons interact with the world, they carry with them an implicit awareness of the basic elements of their histories and anticipated trajectories which at each moment influences both their experience of the present and their deliberations about what to do next. According to this account, persons are able to have the kinds of experiences and engage in the kinds of behaviors they do precisely because they

bring their ongoing life stories to bear on the present and so structure their experience of the world according to an ongoing autobiographical *narrative*. The unity of a single person, according to this view, is the unity of a narrative, and the events that are included in a person's autobiographical narrative are, for that reason, events in *her* life.

This is, of course, only a sketch of a rather complicated position. A full development of the view is beyond the scope of this discussion, but to see its implications for questions of *personal identity* in DID, it will be necessary to have at hand a few more details about the view and the way in which it is motivated. The Narrative Self-Constitution View falls within a broad tradition that employs a particular understanding of what it means to be a person. This understanding defines *personhood* in terms of the kinds of capacities that are characteristic of typical adult humans and are, so far as we know, unique to human existence. These capacities get described differently by different theorists, but crucially they involve the extremely complex kinds of interactions persons are able to have with one another. In particular, there is emphasis on the fact that persons are beings who are subject to norms of thought and action – they can be legitimately criticized for being immoral or irrational in ways that other animals cannot.

The task those who work in this tradition set themselves is that of explaining what relation or relations must be used to define persons in order to capture this fundamental characteristic. It is against this backdrop that the Narrative Self-Constitution View is developed. The idea is that these kinds of judgments are possible because there is a very basic and general conception of what constitutes an intelligible human life that all persons share. There are of course variations of detail over time and among cultures and subcultures, and fine-tuned ideas of intelligibility are variable. There is, however, enough overlap to yield a robust conception of personhood defined in terms of what is held in common – the most basic conditions of intelligibility. The Narrative Self-Constitution View argues that in order to be an entity rightly held to the norms that define persons, one needs first to understand these basic norms and see them as a demand on one's own life. Someone becomes a being capable of moral *agency* or prudential reasoning, and so a being subject to the norms they imply, by understanding what it is to be a moral agent or prudential reasoner and understanding also that he is one. The cognitive and other skills required for *being* a person are the same as those required for *recognizing oneself* as a person among other people. The Narrative Self-Constitution View argues that the form this recognition takes is precisely the structuring of one's life according to an implicit autobiographical narrative which meets the constraints of intelligibility for the biography of a person. This involves both interpreting one's life in terms of these norms of intelligibility and also actively guiding one's life, so far as possible, to meet them.

This requirement implies that personal identity cannot be successfully constituted by just any self-understanding that is basically narrative in form. Someone cannot make any actions and experiences his, simply by telling himself a story in which they are his. Someone may, for instance, sincerely believe that he led the troops at Waterloo on June 18, 1815, but this does not make him Napoleon.

In order to effectively constitute a person, there are two constraints an autobiographical narrative must meet. The first is the "*articulation constraint*." Although the autobiographical *narrative* envisioned in the Narrative Self-Constitution View is mostly implicit, it is required that a person be able to articulate the narrative locally where appropriate. Roughly this means that in the face of a range of questions – e.g., How long have you been here? Where do you live? What do you do for a living? Why did you choose that car? Why did you yell at him like that; what he did seemed reasonable? – a person should have something to say. The answer may sometimes be along the lines of "I don't know, I'm drawing a blank right now," or "I don't know, I make it a point to express my emotions as I feel them and I just don't reflect any further," or even "None of your business." The requirement is not that a person's motives will always be transparent to her or that she must always produce them when asked but rather that she must recognize herself as subject to a certain explanatory burden, at least internally. An individual who does not know where she comes from or if she has a family or a job or why she did what she did *and* is completely unperturbed by this fact is, according to this view, lacking a key feature of personhood.

The second constraint an identity-constituting narrative must meet is the "*reality constraint*." It requires a narrative to conform to the basic contours of accepted facts, where this is understood in a common, everyday sense, and not in any strong metaphysical sense. A narrative that involves getting from Chicago to Paris in under 40 s, for instance, or changing sexes every other day, or being in two places at one time is ruled out. The individual with Napoleonic delusions described earlier would not meet this requirement because he could not give a coherent and realistic narrative that involves undertaking Napoleon's actions. The violation does not lie directly in the fact that he thinks he is Napoleon when he isn't (that would make the view circular), but rather in details, he would need to include in his narrative about what height he is, where he lives, what language he speaks, who he is married to, what year it is now, what year the exploits he is claiming as his own took place, and how long humans live. In order to coherently claim Napoleon's actions as his own, this individual would need to deny accepted matters of fact, thus violating the reality constraint.

In basic outline, then, the Narrative Self-Constitution View says that we constitute our identities as persons by generating and operating with an autobiographical narrative that meets the articulation and reality constraints. There is another way of phrasing the guiding idea behind these constraints that will be helpful in what follows. The Narrative Self-Constitution View holds that personhood requires not only that someone have a conception of herself as subject to the norms of personhood but that her first-personal autobiography must fundamentally cohere with the natural third-person view of her history. There will, of course, be disagreements on points of interpretation – someone may think she is gregarious and funny, while others think she is obnoxious, or she may think that she is full of integrity, while others think she is rationalizing her own bad behavior – because there is usually more than one intelligible way to interpret a history. Small mistakes of detail may also be made from either the first- or third-personal perspective, and these too will

cause minor disagreements. If, however, a person fails to agree with others about the very basic kinds of facts that are part of the *reality constraint*, or about the kinds of explanations for actions that are taken to be minimally intelligible, she will not be able to interact with other people in the ways that characterize personhood.

This provides a broad outline of the Narrative Self-Constitution View. The next section considers what this view says about personal identity in DID patients.

Narrative and DID

At first glance, it may seem as if the Narrative Self-Constitution View would have the implication that there are several distinct persons present in a DID patient. Each alter, after all, seems to have a narrative self-conception of her own, distinct from that of the other alters. Even if there is leakage of skills or some forms of memory, the autobiography with which each alter operates in the world is different. Each has her own sense of how old she is, what she is like, what she has done, who her friends are, what her profession is, and so on. To the extent that we find the normative constraints of narrative self-constitution operating, they are operating *within* an alter.

Things are not quite so straightforward, however, and this is not the whole story the Narrative Self-Constitution View has to tell about these cases. For one thing, it is crucial to note that the individual alter narratives are clearly and dramatically deficient with respects to the norms of autobiographical narrative. This is something of which the patient herself often complains. There are periods of blackout and gaps that cannot be filled in. An alter may suddenly find herself in a strange place with no awareness of how she got there or facing implications of the actions of another alter which are now attributed to her. The kinds of seamless interactions usually taken for granted in daily life are not regularly possible for the alters because they cannot count on the degree of continuity of experience or *agency* that underlie these everyday transactions. The fact that "someone else" has controlled the body to which an alter "returns" after a hiatus makes the standard kind of person-constituting narrative impossible.

This does not, by itself, always mean that the alters are not individual persons. As explained above, an identity-constituting *narrative* is not required to be perfectly accurate or complete; all that is required is that the individual recognize the norms of accuracy or completeness as legitimate goals. This is something that at least some of alters in most cases of DID clearly do; they are in fact deeply distraught by the disruptions of narrative continuity and lack of coherence in their lives. Nevertheless, the narrative lacunae within the individual alter narratives are usually sufficiently profound that they interfere markedly with the ability to engage in the kinds of interactions that characterize and define personhood in the sense at issue here. This is one reason the condition is considered pathological.

There is, moreover, another feature of DID which is directly relevant to questions of *personal identity*. That is the fact that there is another perspective, external to that of any of the individual alters, from which all of the alters are most naturally construed to be parts of a single (albeit atypical and disjointed) life narrative. Whatever the outcome of debates on the etiology of the disorder, from the therapeutic perspective the story is one in which a single person presents as psychologically fragmented. On the standard picture, it is a story of someone with a certain kind of susceptibility who is traumatized at a crucial life stage and protects herself by compartmentalizing her psychological life and dissociating from painful stimuli. The various alters are seen as pieces of the single person who take on different emotional jobs, and the hope is that the parts can be reintegrated so that the person can be whole again. This is equally clear, if not clearer, when the disorder is seen to be an artifact of suggestion. Here the narrative is one of a single vulnerable and impressionable person whose psychological life becomes disordered as a result of therapeutic suggestion or media hype.

It is worth noting as something of an aside that questions of etiology interact with the Narrative Self-Constitution View in many interesting ways. Among other things at issue in the debate over this aspect of DID is the question of whether those who claim recovered memories of abuse are meeting the *reality constraint*. The traditional view assumes that in most cases they are, while skeptics about the condition assume that most recovered memories are confabulation. There are many fascinating questions to be considered about the implications of this dispute, and these are worthy of careful study. They are, however, somewhat orthogonal to the issues currently under consideration. Whichever position one takes on the question of the origins of the phenomena associated with DID, there is a natural third-personal narrative of a single person who suffers some kind of psychological event that leads to symptoms of disorder and division. To the extent that the therapeutic goal is the production of a unified and integrated whole, this is a story of a single person with a psychiatric disorder rather than of several different people who happen to share a body.

The situation in DID thus seems to be one in which there is a sharp disconnect between the first-personal narrative and the third-personal narrative. The alters understand themselves as distinct individuals with their own biographies, whereas from the outside a single biography of trauma (or suggestion) leading to psychological disorder is more natural. The division is not, of course, as clear-cut as has been suggested so far. First-personal and third-personal perspectives always infect one another; that is the very nature of *personhood*. Those who interact with DID patients will likely not be able to *simply* see the patient as a single individual, since different alters must be interacted with as if they were distinct persons. And while the patient might not *experience* herself as a single person in the standard way, given a diagnosis of DID, she (at least some of the alters) may well *see* herself as a single individual with a rare disorder. We thus need a more complex description of the kind of disconnect between first- and third-personal perspectives in DID to fully understand how the Narrative Self-Constitution View analyzes this phenomenon.

It will be useful, for these purposes, to look at some similarities and dissimilarities between DID and non-pathological multiplicity. It has long been recognized in both philosophy and psychology that it is standard for humans to have psychological lives that are fragmented and compartmentalized in various ways. We experience the world differently (and so behave differently) in different contexts, in different moods, and at different life phases. According to the Narrative Self-Constitution View, the accomplishment of constituting oneself as a person involves recognizing this fact and, in a particular way, compensating for it. In non-pathological cases, persons recognize the different perspectives they experience as one and all their own and so recognize themselves as bearing some responsibility for making them cohere, or at least coexist. This, in turn, leads to the phenomenological and behavioral changes characteristic of narrative self-understanding.

An example may help. In Elliot's *Middlemarch* (1985), the narrator describes the state of mind of Dr. Lydgate during his impetuous and ill-fated plan to propose to an actress with whom he has fallen madly in love: "He knew that this was like the sudden impulse of a madman – incongruous with his habitual foibles. No matter! It was the one thing which he was resolved to do. He had two selves within him apparently, and they must learn to accommodate each other and bear reciprocal impediments." (p. 182) Lydgate has "two selves" – the madly infatuated lover and the responsible doctor – but he knows that he has these two selves and experiences them both as *him*self. He is the overarching whole within which these selves reside, and their conflict is internal to *him*. Since these selves are both part of Lydgate they need to "learn to accommodate each other and bear reciprocal impediments." This means, among other things, that the desires, proclivities, and goals of each put constraints on those of the other. Lydgate does not simply suppress the pursuit of the lover nor does he allow it to get to the point where the rational doctor is crowded out entirely. Even in the throes of passion, he recognizes himself as a single, complex being and works to create a coherent *narrative* that involves both of these selves.

That Lydgate thinks of himself in this way shows that he has internalized in a particularly deep way the third-person perspective which sees him as a single individual. Others see him as a single person, despite the very different inner experiences he has, and this leads him to understand himself as a single person in a way that changes and constrains the vicissitudes of his inner life. The phenomenology of his infatuation and his responses to it are all conditioned by the fact that it is experienced as part of a larger life which includes both the lover and the responsible doctor. The effect here is much the same as that described earlier in which the experience of disembarking from an airplane is conditioned by the larger story within which it occurs. It is this kind of deep internalization that is missing in DID patients. They may (or may not) come to *think* of themselves as single individuals, but they do not *experience* themselves as such; the attitudes, experiences, and actions of the alters do not interact with and change one another. It is possible to see them as parts of a single life story but only, as it were, from an external perspective.

Applications and Insights

The previous section provided an overview of the way in which the Narrative Self-Constitution View can be used to analyze questions of identity in cases of DID, but it has not yet provided an answer to the original question about the number of persons encountered in encountering a DID patient. In the end the Narrative Self-Constitution View delivers no straightforward judgments about this question. On the one hand, each alter seems to have something like a personal *narrative*, and this suggests that there are multiple persons alternating control of a single body. The narratives of the alters are, however, necessarily gappy and deficient, and from the third-person perspective, there is a natural single-life narrative to tell about an individual with a fragmented psyche. This suggests that there is a single person present who is suffering from a particular psychopathology. The third-personal narrative is not a standard life story for a person, however, since it involves an anomalous degree of fragmentation and incoherence. Because of this, and because the first-personal and third-personal narratives do not mesh in the way that is typical for human lives, neither the individual alters nor the human in whom they alternate present an unproblematic life narrative. From yet another perspective, then, it seems as if there is *no* whole person present.

The Narrative Self-Constitution View thus helps to clarify exactly why it is so difficult to say how many persons are present in such cases. It also illuminates the role of embodiment in personal identity. The difficulties that arise with the narratives of DID patients stem from the fact that we expect a certain kind of narrative coherence within a single *human* life. There are good reasons for this constraint. Personhood, as understood here, is defined in terms of the ability to interact with others in the particular complex ways that typify human life. These interactions usually require us to be able to quickly identify those with whom we are interacting and to reliably make certain kinds of broad predictions about what information they are likely to have and how they are likely to behave. Since persons act through and with their bodies, moreover, they need to be able to count on a large degree of regular control over what their bodies do if they are to be effective agents. Embodiment must be taken into account when thinking about *personhood* and *personal identity*, and this is something that those who define persons as subjects and agents do not always fully appreciate. This does not mean that personal identity can or should be defined directly in terms of biological continuities, however (e.g., Olson 1997). Merely being a single human being does not make one a single person. Personal unity must be accomplished through the development of a unified self-narrative that shapes one’s experience and behavior; the role of embodiment is to delimit what can and must be included in such a narrative if it is to be truly person-constituting.

While the Narrative Self-Constitution View does not provide a simple answer to the question of how many persons are present when we encounter a DID patient it can, nevertheless, provide a rich and valuable theoretical apparatus which can be fruitfully applied to the many practical puzzles this disorder engenders. As an example, consider the question of the assignment of responsibility. Questions

frequently arise about whether one alter should be held accountable for actions taken by another. The force of these questions can be seen in legal quandaries involving criminal cases in which the defendant is understood to have a legitimate diagnosis of DID. Behnke (1997) describes three different legal approaches that have been taken to this problem. One holds that individuals with DID are generally not responsible for their crimes because the alters are enough like persons to be treated like persons. If the alter who committed a criminal act is sufficiently person-like to be held responsible, according to this position, then alters who have no knowledge of or role in the crime must also be treated as persons, and it is unacceptable to punish them for what they did not do and could not control. At the other extreme is the view, defended by Behnke himself, that we must distinguish between "person" and "personality" and recognize that it is persons rather than personalities that act. For this reason, he thinks that defendants with DID generally can be held responsible for criminal actions. While it is true that we are confronted with divided personalities, he says, there is only one *person* present, and there is little doubt in most cases that that person is capable of taking action for which she is rightly held culpable. There may be, for instance, extensive planning or calculated attempts to evade detection. In between these poles is the approach that takes the host, or dominant, personality to be the individual relevant to the assessment of criminal responsibility; if the host personality was not involved in the planning and execution of the crime, the view says, the individual should not be held culpable.

Common to all three of these approaches is an attempt to locate the person or persons involved in the crime. The first extreme sees each alter as a person, the second sees the human as a whole as the person, and the middle position sees the host personality as the person. If the Narrative Self-Constitution View is correct, however, this general strategy is mistaken. The alters *are* person-like in respects that are typically connected with *agency* and responsibility, and in some ways, perhaps, the host personality more so than any of the others. But actions *are* undertaken through human bodies, and the human body on trial is the same one allegedly involved in the crime. Rather than trying to resolve these competing pulls, the Narrative Self-Constitution View suggests that we recognize DID as an impairment of *personhoo*d that makes it impossible to find a paradigmatic moral agent. The question we thus need to ask is exactly what impact this particular kind of impairment of personhood has on the relevant forms of agency and hence on the assessment of moral or criminal responsibility.

This approach is in some ways like another defense sometimes offered in criminal trials of DID patients which argues that such patients should not be held criminally responsible for their actions either because the action is involuntary or because of defects in mens rea. This approach basically sees the defendant as a single person who is not responsible for her actions. This position also does not seem quite right either, however. As Behnke (1997) puts it, "the problem for criminal courts is that neither involuntariness nor insanity speaks clearly to the central feature of MPD [DID]: dividedness" (p. 392). The kind of impairment of *personhood* found in *DID* does not undermine agency in the way that brainwashing

or psychosis would. At the time of any particular action, the individual actor is likely to be in control of her behavior and to be able to distinguish right from wrong and reality from fantasy to the extent usually necessary for criminal culpability. The impairment is more subtle; it has to do with the inability to put together a coherent inner narrative that corresponds to the history of a single human animal.

Recognizing this problem does not offer an immediate solution to the question of how to assess moral or legal responsibility in cases of DID, but it does show us where to look to think productively about these cases. The puzzle arises because it looks for all the world as if there is a minimally competent agent acting locally at the time a criminal action is taken, but there nevertheless seems to be a problem of agency because of the way in which agency *at a time* is impacted by failures of coherence *over time*. The question of the connection between agency at a time and coherence over time has received some attention from philosophers, but not as much as it should have. The Narrative Self-Constitution View offers a particularly useful framework for investigating these issues by showing the ways in which diachronic and synchronic aspects of personhood and agency are interconnected, and so how failures of coherence over time can affect agency at a time. This is just what is needed to determine agency and responsibility in these cases. There is much work to be done to develop the understanding we need to answer these questions fully, but the Narrative Self-Constitution View offers a path to getting that work done.

Conclusions and Future Directions

DID troubles our ordinary way of understanding and individuating persons. While it is tempting to try to ascertain how many persons are present in a case of DID, it is more fruitful instead to recognize that there is no straightforward answer to this question and to find a theoretical framework for understanding the different ways in which this disorder interferes with paradigmatic personhood, drawing the practical implications of and potential remedies for such impairment. The *narrative* approach promises to provide such a framework.

Cross-References

► Consciousness and Agency
► Ethics in Psychiatry
► Extended Mind and Identity
► Free Will, Agency, and the Cultural, Reflexive Brain
► Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System

References

Baylis, F. (2012). The self in situ: A relational account of personal identity. In J. Downie & J. Llewellyn (Eds.), *Relational theory and health law and policy* (pp. 109–131). Vancouver/ Toronto: UBC Press.

Behnke, S. H. (1997). Assessing criminal responsibility of individuals with multiple personality disorder: Legal cases, legal theory. *The Journal of the American Academy of Psychiatry and the Law, 25*(3), 391–399.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Coons, P. M. (1988). Psychophysiologic aspects of multiple personality disorder: A review. *Dissociation, 1*(1), 47–53.

Dennett, D. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, & D. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (pp. 103–115). Hillsdale: Erlbaum.

Dorahy, M. J. (2001). Dissociative identity disorder and memory dysfunction: The current state of experimental research and its future directions. *Clinical Psychology Review, 21*(5), 771–795.

Eich, E., Macaulay, D., Loewenstein, R. J., & Dihle, P. H. (1997). Memory, amnesia, and dissociative identity disorder. *Psychological Science, 8*(6), 417–422.

Eliot, G. (1985). *Middlemarch*. Middlesex: Penguin Classics.

Goldie, P. (2012). *The mess inside: Narrative, emotion & the mind*. Oxford: Oxford University Press.

Loewenstein, R. J., & Putnam, F. W. (1990). The clinical phenomenology of males with MPD: A report of 21 cases. *Dissociation, 3*(3), 135–143.

Loftus, E. F. (1993). The reality of repressed memories. *The American Psychologist, 48*, 518–537.

Merckelbach, H., Devilly, G. J., & Rassin, E. (2002). Alters in dissociative identity disorder: Metaphors or genuine entities? *Clinical Psychology Review, 22*, 481–497.

Olson, E. (1997). *The human animal: Personal identity without psychology*. Oxford: Oxford University Press.

Piper, A., & Merskey, H. (2004). The persistence of folly: A critical examination of dissociative identity disorder. Part I. The excesses of an improbable concept. *Canadian Journal of Psychiatry, 49*(9), 592–600.

Reinders, A. A., Nijenhuis, E. R., Quak, J., Korf, J., Haaksma, J., Paans, A. M., Willemsen, A. T., & den Boer, J. A. (2006). Psychobiological characteristics of dissociative identity disorder: A symptom provocation study. *Biological Psychiatry, 60*(7), 730–740.

Ricoeur, P. (1992). *Oneself as Another*. (trans.) Kathleen Blarney. Chicago: University of Chicago Press.

Schechtman, M. (1996). *The constitution of selves*. Ithaca: Cornell University Press.

Schechtman, M. (2011). The narrative self in Shaun Gallagher. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 394–418). Oxford: Oxford University Press.

Taylor, C. (1989). *Sources of the self*. Cambridge, MA: Harvard University Press.

Impact of Brain Interventions on Personal Identity

25

Thorsten Galert

Contents

Abstract

Some of the main ethical concerns with brain interventions are related to psychological changes that may alter persons in morally significant ways. This chapter reviews neuroethical attempts at assessing brain interventions by describing them as having an impact on the identity of persons. By discussing clinical case reports involving deep brain stimulation in the treatment of Parkinson's disease, different meanings attributed to the concept of personal identity in the bioethics literature are disentangled. The concept of forensic personhood is introduced to show that some possible effects of brain interventions are unambiguously detrimental in that they end the affected person's existence in an important sense. The same holds true for changes in personal identity where a person has changed to such an extent that, eventually, one does not seem to be faced with a person with different characteristics, but with an altogether different person. As well, narrative accounts of personal identity are

T. Galert
German Reference Centre for Ethics in the Life Sciences, Bonn, Germany
e-mail: galert@drze.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_12,

discussed in some detail as they have multiple normative implications for the assessment of the psychological effects of brain interventions, ranging from legal and ethical to clinical and methodological consequences.

Introduction

Some of the main ethical concerns with brain interventions are related to psychological changes that may alter persons in morally significant ways. While the risk of severe psychological adverse effects is not unique to brain interventions (cf. Lipsman et al. 2009, p. 376), such interventions are particularly likely to provoke fears about radical changes to a person's frame of mind. This is because these interventions seem to have a direct impact on the biological substrate of personal identity. It is beyond the scope of this chapter to review the rich philosophical debate on whether it is necessary to refer to neurobiology to effectively address the question of whether a person remains the same person over time in spite of certain psychological changes. Rather, the focus here is on distinguishing different ways in which a person's existence may be threatened by brain interventions. As only some of these threats can properly be described as involving genuine changes of personal identity, these conceptual clarifications should help disentangle the different meanings attributed to the concept of personal identity in the bioethics literature. Subsequently, a particular strand of theorizing that addresses issues of personal identity in narrative terms will be introduced in more detail. Since narrative approaches to identity appear particularly applicable to clinical cases involving untoward consequences of brain interventions, it should be clear why they came to play such a prominent role in relevant neuroethical debates.

The conceptual analysis will be advanced by discussing clinical case reports involving the application of deep brain stimulation (DBS) in the treatment of Parkinson's disease (PD) (cf. ▶ Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"). The reason for focusing on DBS in PD is not that this neurotechnical intervention is particularly threatening for personal identity nor that it would involve any special risks for the integrity of persons. Rather, this focus is motivated by a desire to limit the empirical complexities, as the conceptual matters are complicated enough. Issues of personal identity are regularly brought up in the ethical assessment of neurosurgical interventions in general (cf. ▶ Chap. 59, "Ethics in Neurosurgery") and invasive neurotechnologies like DBS in particular (cf. ▶ Chap. 34, "Ethical Implications of Brain Stimulation").

In the early days of neuroethics, neural grafting (cf. ▶ Chap. 52, "Ethical Implications of Cell and Gene Therapy"), e.g., the transplantation of small amounts of fetal brain tissue for the treatment of PD, was the most frequently discussed threat to personal identity (Boer 1994). In recent years, DBS has replaced neural grafting as the prime example of a brain intervention that might cause changes in personal identity. This shift can be explained by the fact that invasive brain stimulation techniques

have made significant progress in a short period of time, moving from small case studies to clinical practice. DBS has not only been established as an effective and reasonably safe therapy for several neurological disorders, but it is also increasingly applied in the treatment of severe psychiatric disorders (cf. ► Chap. 39, "Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review"). With the increasing use of DBS, anecdotal evidence has accumulated confirming the potential for deep and sometimes disturbing psychological impacts. Some of this evidence will be reviewed carefully before turning to the conceptual question of how the observed changes can be most appropriately described.

Two Clinical Cases of DBS in PD

Case report 1: In a Dutch case, described by Albert Leentjens and colleagues (2004), a 62-year-old male patient developed a manic state after 3 years of stimulation of the subthalamic nucleus (STN) for severe PD. His psychiatric symptoms included megalomania and chaotic behavior that had resulted in serious financial debts. The patient did not respond to treatment with a mood stabilizer and so was admitted to a psychiatric hospital where he was deemed incompetent. By adjusting the stimulation parameters, it was possible to restore his decisional capacity. Unfortunately, this improvement resulted in a worsening of his motor symptoms. It was impossible to strike a satisfying balance between alleviating the psychiatric and the neurological symptoms, and so the patient ultimately was presented with a terrible choice: He could be admitted to a nursing home, physically disabled, but in good mental health, or he could be committed to a psychiatric ward because of his mania, but with good motor function. With the stimulator turned off, and the patient deemed competent, he chose to be admitted to a psychiatric hospital.

Case report 2: A much less dramatic, but no less important, case involving DBS for the treatment of PD has been reported by Michael Schüpbach and colleagues (2006). A 48-year-old male accountant, whose PD symptoms were successfully alleviated by DBS, experienced a grave marital conflict in the year following the initiation of DBS. Before receiving DBS, the patient was dependent on his wife for all tasks of daily living. Following DBS, he regained self-confidence and this resulted in marital strife. The authors describe the ensuing conflict by quoting the patient and his wife:

> "I want to recover my social standing and establish new relationships outside my couple. During all these years of illness, I was asleep. Now I am stimulated, stimulated to lead a different life." Confronted with the radical change in her husband's behavior, his wife became depressed: "Ever since the operation, I feel lost. Before, when he was sick, we were a perfect couple. Now, he wants to live the life of a young man, go out, meet new people, all of that is intolerable! I would rather he be like he was before, always nice and docile!" The patient persisted in his desire for change: "All these years I allowed myself to be carried like a child, because I didn't have the means to fight. That period is over, I want to get back the position I left open. I am going to take my life in hand, my life before PD." (Schüpbach et al. 2006, p. 1812)

These two case reports illustrate the wide range of impacts that DBS can have. One important difference between the two cases concerns the causal link between the intervention and the outcome. The problematic consequences experienced by the patient in the first case can be causally attributed to brain stimulation in a more immediate way than in the second case. The manic condition of the first patient is quite obviously a direct effect of DBS, as it can reliably be altered by initiating or stopping the stimulation. With the second patient, turning off the stimulator could return him to his previous neurological state, but this probably would not resolve the conflict with his wife (which makes the psychological consequences of DBS appear problematic in the first place). The marital problem has been caused by brain stimulation in some sense, but for this cause to take effect, a whole range of mediating psychosocial factors had to fall into place. As these factors can’t be made to disappear just by stopping DBS, the problematic sequelae may well persist even if the stimulation is stopped.

A second important difference between the two clinical examples concerns the different perceptions of the resulting change(s). As regards the first patient, the psychological change induced by DBS is clearly to be considered an *adverse* effect of stimulation (even if the patient enjoyed himself in his manic state – i.e., if he did not personally experience his state of mind or chaotic behavior as problematic). With the mania the patient loses something that is so universally and strongly evaluated as a “good” that the resulting state will be considered detrimental irrespective of its possible appreciation by the patient (cf. Schechtman 2009, p. 70). This “good,” of course, is the capacity to decide what is in ones best interest and, at the same time, what is acceptable for others in the light of widely shared normative standards. In contrast, the second patient’s rekindled lust for life does not seem to bear the hallmark of psychiatric abnormality. The patient’s appreciation of his regained independence seems perfectly understandable and healthy. In fact, if the patient’s psychological transformation did not cause so much discomfort to his wife, one would have no reason for moral suspicion regarding the effects of brain stimulation in this case.

Ways of Questioning the Identity of Persons

Is it appropriate in either of the two cases to describe the changes as changes in personal identity? Walter Glannon analyzes the first case (involving the Dutch PD patient) in two papers (Glannon 2009, 2010) focusing on patient autonomy in general, and more particularly on whether a patient can give valid consent to a treatment that deprives him of the capacity to revise this decision in the further course of treatment. In passing, Glannon addresses the suspicion that the patient is undergoing identity change when switching into the manic state. According to Glannon, this question must be answered in the negative if one presupposes a concept of *numerical* identity. Glannon maintains that continuity in “a minimal set of physical and psychological properties” suffices to make sure that an individual remains the person he or she is in the sense of numerical identity (2009, p. 291).

If, on the other hand, one presupposes a concept of *narrative* identity, then the crucial question is whether the disruption in the patient's personality caused by DBS, as represented in his "distinctive autobiography," is severe enough to claim that, under its influence, the patient becomes a different person (ibid.).

The Forensic Account of Personhood

The basic assumptions of narrative views of personal identity will be addressed in more detail below. Before doing so, however, it is worth considering a third possible description of the change induced by DBS in the first case. In this case, the transition into the manic state is characterized by a severe loss of decisional capacity – which is not just any kind of personality trait. This loss renders the patient legally incompetent to the point that institutionalization is warranted. Under the influence of DBS, the patient can still act, but his actions are no longer autonomous in the important sense that he can no longer be held fully responsible for them. According to a long-standing tradition in philosophy, going back to John Locke, this amounts to saying that DBS deprives the patient of an essential feature of personhood. Locke considers "person" to be a "forensic term" whose primary function consists in "appropriating actions and their merit" (Locke 1975 [1689]: §26). Hence, a human being who, for whatever reason, lacks the requisite psychological capacities to engage in moral and legal practices like giving promises or entering into other sorts of binding commitments cannot be regarded as a person in the proper sense (cf., e.g., Merkel et al. (2007: 5.3.2) for a detailed account of the cognitive, emotional, and motivational requirements that such a view of personhood entails). Following Locke, Marya Schechtman proposes to distinguish a further possible sense of identity, namely, "forensic personal identity" (2009, p. 68). According to Schechtman, the "judgment that someone is not (or not fully) a forensic person relieves him of certain kinds of responsibilities and commitments" (2009, p. 69). What such a judgment does not entail, however, is the denial of rights (in particular the right to respectful treatment) that need to be respected in relation to any human being, regardless of whether he/she is a person.

This account of personhood in terms of moral agency necessitates the distinction between a loss of personhood on the one hand and genuine cases of (forensic) identity change on the other hand. Schechtman exemplifies the former possibility by referring to the gradual fading away of forensic personhood during the progression of dementias such as Alzheimer's disease which can, eventually, lead to a state in which a patient may no longer be held accountable for his actions (2009, p. 71). Schechtman's discussion of the second possibility, personal identity change, is based on a (hypothetical) clinical case involving DBS for a Mr. Garrison for the treatment of advanced PD – a case that bears many similarities to our second case of the reinvigorated husband.

In Mr. Garrison's case, DBS has the wanted side effect of relieving him of apathy syndrome which he developed during PD progression (2009, p. 77).

Mr. Garrison's lifted spirits provoke marital problems. Under the influence of DBS, he pursues interests and develops passions that, at least as judged by his wife, do not fit the person that Mr. Garrison was before brain stimulation. For instance, he starts spending money on charitable causes, thereby threatening plans that he and his wife had made in caring for their retirement. Also, under the influence of DBS, he is outgoing and gregarious, where once he was shy and introverted. Despite the considerable discontinuity in Mr. Garrison's personality before and after DBS, according to Schechtman it would be wrong for his wife to acquit herself of her duties as a spouse by making claims to the effect that she did not marry the person that Mr. Garrison is under the influence of DBS. Schechtman considers the fact that Mr. Garrison's radical change of attitudes can at least to some extent be attributed to the influence of DBS as a kind of mitigating factor, because "it does not seem entirely fair to hold Mr. Garrison to the same standards of consistency that we would expect from someone who had not had such treatment. It does not seem obvious that he should be held responsible for commitments made before the personality change, and to this extent it seems right to treat him as a new forensic person" (Schechtman 2009, p. 78).

Mr. Garrison's case illustrates how recognizing a change in forensic personal identity can be complex and ambiguous. At least with regard to the practical commitments of a married couple, any single decision as to whether such an identity change occurred seems to be very much context dependent and, ultimately, open to negotiation. When the stakes are higher, however, as for instance in cases of potential legal liability, less ambiguity will be tolerated in deciding issues of forensic personal identity. For example, if Mr. Garrison committed murder before receiving DBS, in all likelihood the psychological discontinuities brought about by DBS would not be judged severe enough to excuse the stimulated Mr. Garrison from criminal liability for a crime that could not be attributed to any other human being but him. Were he careless enough, for instance, to testify that he perfectly understood why the person who used to reside in his body before DBS murdered the victim and that he himself would do the same if he had the chance, then any jury would probably feel entitled to conclude that there is sufficient personal continuity between the accused person and the person who committed the crime to justify Mr. Garrison's conviction in spite of his alleged identity change. On the other hand, matters might be different if Mr. Garrison had no memory of the crime.

Reinhard Merkel and coauthors (2007, p. 270) consider cases of dissociative amnesia as the most clear-cut examples of personal identity change. If a person loses all his autobiographical memories, as can happen with complete retrograde amnesia, it may seem appropriate to say that, in a significant sense, the earlier person's existence has come to an end. If the individual who suffers from amnesia still fulfills the criteria of personhood, then he may legitimately claim to be a new person who has to start his life from scratch. Merkel and coauthors admit, however, that even in such extreme cases of psychological discontinuity, a person would not necessarily be acknowledged as a *new* person, "but rather as a person with a certain form of dissociative disorder" (2007, p. 270).

Normative Implications of the Concept of Forensic Personhood

Any answer to the question of whether a change in personal identity may result from a brain intervention obviously has to presuppose a theory of personhood. For without knowing what persons are, it is impossible to say under which circumstances a person comes to an end and another emerges. With the concept of forensic personhood, persons essentially are moral subjects who can take responsibility for their actions. Using this conceptual framework, the special tragedy of the Dutch patient lies in the fact that by deciding to receive DBS he is deliberately entering a state in which his accountability is reduced to such an extent that he can no longer be recognized as a forensic person. This case is vexing in that the patient under the influence of DBS in most respects still behaves like a person. It is certainly easier to acknowledge the loss of personhood in human beings who are in long-lasting states of unconsciousness. These difficulties notwithstanding, if moral agency is at the heart of personhood, then the person who the Dutch patient used to be prior to receiving DBS has ceased to exist. Only by making a clear distinction between persons and human beings can theories of forensic personhood also leave room for the possibility that the human being who used to be recognized as a particular person before an intervention into her brain continues to live as a different person. While indisputable cases of forensic identity change are hard to find, the pivotal issues are whether and to what extent a person after a brain intervention is both willing and able to take responsibility for previous actions and live up to the commitments and consequences that follow from those actions. Issues concerning personal identity change after brain interventions will only be raised if a patient severely fails to fulfill the expectations of others related to the commitments of the person he used to be before an intervention into his brain.

It is useful to make a clear conceptual distinction between cases in which a brain intervention like DBS leads to (a) the loss of personhood, (b) a change of (forensic) personal identity, and (c) other more or less profound psychological sequelae, so as to offer a nuanced normative assessment of these different classes of consequences. Any medical intervention that regularly leads to consequences of either type (a) or (b) merits the highest ethical scrutiny as in these instances the person’s existence comes to an end – an event that is sometimes called “psychological death.” Nevertheless, such interventions can be justified as legitimate treatment decisions if foregoing them might lead to similarly grave consequences. The physicians who turned the stimulator back on after the Dutch patient consented to the continuation of DBS seem ethically justified in what they did, only because it was the patient himself who made the decision. It would be impossible to decide for someone else whether it is best to be of sound mind in an immobilized body or to retain motor control over ones actions at the cost of losing the ability to take full responsibility for those actions.

Because of the normative salience that any impact on personhood and personal identity acquires in a framework of forensic personhood, it would also seem justified to require that it should never be among the goals of any intervention in the brain to bring about consequences of type (a) or (b). Such consequences can be

envisaged as grave risks associated with desperate treatment decisions, but they should never be viewed as possible benefits. Compared to these strong normative claims, we are entering the realm of ethical ambiguity when assessing the cluster of possible psychological changes provisionally lumped together under (c). In order to establish further meaningful normative distinctions for the assessment of possible psychological effects of brain interventions, we need to review the conceptual resources of narrative theories of personal identity.

Narrative Approaches to Personal Identity

As noted above, philosophers who champion a narrative approach to personal identity tend to draw a sharp distinction between narrative and numerical identity (cf. also DeGrazia 2005, p. 78 and Focquaert 2009, p. 2). To properly grasp this distinction, one has to understand that two quite different concepts of a person's "identity" are in play. A person's numerical identity is at stake when the task at hand is to make sure that a person at one point in time is the same person at another point in time. Here "identity" is taken to refer to the logical relation of identity (cf. Merkel et al. 2007, p. 195). Theories of numerical personal identity thereby answer the *reidentification* question by specifying conditions under which a person can be correctly reidentified as the same person in spite of her having changed over time in various ways. According to Schechtman (1996), who developed a narrative approach to personal identity that is particularly influential in bioethics, an alternative perspective on a person's identity is provided by the *characterization* question which is asking "which beliefs, values, desires, and other psychological features make someone the person she is" (Schechtman 1996, p. 2). The characterization question addresses a person's identity in roughly the same meaning as implied in talking about a person suffering an "identity crisis" (ibid., p. 74), giving rise to questions like "Who am I really?" A person's identity, in this sense, can be described by a set of characteristics. On this view, "identity change" is shorthand for changes in a person's characteristics that are profound and/or pervasive enough to justify the claim that the person is no longer the same in an important sense in spite of his numerical identity being preserved. The characterization question offers a sound way of interpreting sentences such as "this person has changed her identity" or "that person lost his identity" – interpretations that do not make sense if personal identity is understood numerically. From the point of view of reidentification, talking about a change of personal identity implies that a person has undergone a process of change to a degree that, eventually, we are not faced with a person with different characteristics, but with an altogether different person. Taken literally, however, the sentence "this person has changed his identity" presupposes one person's continued existence, albeit with different identities at different points in time.

Viewed in this way, it is misleading to contrast numerical and narrative identity, because the former concept of personal identity is more fundamental than the latter. Rather than providing an independent perspective on what a person's identity is, the

characterization question presupposes an answer on the reidentification question (cf. Merkel et al. 2007, p. 193; DeGrazia 2005, p. 114). Before making meaningful intrapersonal comparisons of a particular person's characteristics at different points in time, one needs to know that the person remains the same in the numerical sense over the whole period of comparison. Thus, it is somewhat inadequate to talk about "narrative identity" next to "numerical identity," because this way of putting things conceals that the two concepts are operating on different levels. Rather than viewing narrative accounts of personal identity as dealing with a special kind of identity, it may be more helpful to describe them as representing a particular approach to questions related to personal identity. While most proponents of this approach restrict their narrative theorizing to the characterization question, Merkel and his coauthors (2007: 5.4.5) propose a narrative approach to personal identity that is supposed to provide an answer to the reidentification question as well. They avoid the term "narrative identity" altogether, instead calling their answer to the characterization question a narrative account of *personality* (2007: 5.4.4). This seems like an interesting suggestion in view of the fact that, as will be shown subsequently (sub 4.2), the assessment of personality changes is a particularly important area of application for narrative accounts of personal identity.

Basic Assumptions of Narrative Accounts of Personal Identity

Just what is "narrative" about narrative approaches to personal identity (cf. also the ► Chap. 24, "Dissociative Identity Disorder and Narrative")? This is not the place to delve deeply into the dazzling diversity of narrative approaches to the "self" advanced by philosophers, sociologists, and social and developmental psychologists (cf. Schechtman 2011 for a review). Instead, the focus here is on common features of narrative approaches to personal identity that surface in discussions of neuroethical issues. Common to such accounts is the idea that persons or selves are *constituted* by the stories they tell about who they are. That is, listening to autobiographical narratives is not just an epistemic strategy to find out who a person is. A person's identity does not simply *express* itself in a self-narrative, rather it is *formed* in the process of autobiographical storytelling. Persons are making sense of themselves by structuring their lives in narrative ways or, as Schechtman puts it, "selves are inherently narrative entities" (2011, p. 395).

Narrative accounts of personal identity need to come to terms with the fact that self-narratives are fraught with mistakes. The stories people tell about who they are include both inadvertently erroneous and deliberately deceptive statements. Hence, such stories do not reveal who a person really is but (at best) only how she conceives of herself, i.e., her self-concept. Schechtman attempts to bridge the epistemic gap between a person's self-concept and any trustworthy account of who the person really is by introducing the *reality constraint*. According to this constraint only those parts of a person's self-narrative which "fundamentally cohere with reality" will be recognized as identity constituting (Schechtman 1996, p. 119). Other proponents of narrative accounts of identity offer different solutions to the

problem of (self-)deceptive self-narration, but they all agree that some kind of reality check for autobiographical storytelling is required (DeGrazia 2005, p. 85; Merkel et al. 2007: 5.4.4).

A further fact that may seem to challenge the basic idea of narrative approaches to personal identity is that people differ widely in their propensity to tell stories about themselves. Some people just do not like to spend many words on characterizing themselves. Should they be viewed as having "no (narrative) identity" or at least as having "no well-developed identity"? – Schechtman avoids implausible conclusions of this kind by pointing out that a person's self-narrative is not only articulated in stories she explicitly tells to others, rather: "Much of our self-narration is expressed in the way we think, the way we live, and the kinds of explanations we feel called upon to give others" (Schechtman 2011, p. 407). The fact that people do regularly demand explanations from each other with regard to the way their behavior may or may not fit with their perceived self-concepts lends credence to the *articulation constraint*. This constraint requires that, at least when prompted to do so, a person should be able to "account for her actions and experiences by showing how they are a part of an intelligible life story with a comprehensible and well-drawn subject as its protagonist" (Schechtman 1996, p. 114). Alternatively, the articulation challenge can be met by highlighting the social interplay in which self-narratives are formed. After all, whenever self-narratives are actually told to others, the listeners are likely to play an active part in revising the narrative by adding their own material and perspective. In view of the social contribution to self-narration, and the presumed fact that our self-concepts are revealed in the way we act, not much explicit storytelling seems to be required on the part of a person for the constitution of a narrative self. Merkel and coauthors confirm this point by claiming: "even the most reticent person, who remains imperturbably close-mouthed whenever others approach her with their unbidden characterisations, cannot avoid to confirm or refute the stories they tell about her through her deeds" (2007, p. 258).

The fact that self-narratives gain and change their form and content over time in an ongoing process of social exchange has obvious ramifications for the reality constraint as well. The listeners of self-narratives not only help in articulating the story of who a person is, they also keep an eye on the storyteller staying true to the facts. Beyond such basic considerations, however, the extent to which the social nature of self-constitutive storytelling is brought to bear differs significantly in the literature on narrative personal identity. For example, Françoise Baylis' "relational account of personal identity" offers an example of a narrative approach to identity that places special emphasis not only on the social, but also the cultural, political, and historical embeddedness of self-narratives. As an alternative to the *reality constraint* and the *articulation constraint* introduced by Schechtman, Baylis introduces the *equilibrium constraint*. According to this constraint, "the identity-constituting narrative is the narrative that effectively balances how a person sees and understands herself, with how others see and understand her" (Baylis 2011).

A further widely shared feature of narrative approaches to personal identity is that, at some point in their theorizing, they adopt a *principle of coherence*.

For example, Merkel and coauthors not only require that self-constituting stories, which purport to describe who a person really is (or "what her personality is like," in the terminology of Merkel et al.), should be "extrinsically" consistent with the facts but also that they "intrinsically" fulfill minimal requirements of coherence that any narrative with a claim to truth needs to meet (2007, p. 261). Similarly, Schechtman demands a high degree of coherence of narratives if they should be recognized as identity constituting, that is, such stories should obey to comparatively "conservative" standards of narrative form (1996, p. 98). Thus, even if narrative theorists need not require that a self-constituting narrative should have an overarching theme or purpose, they may still want the storyteller to at least try to make some sense of the story of his life. In assuming a narrative form, the events of one's life are placed in a comprehensible sequence. Moreover, in arranging the story line, not only the past is taken into account, but also plans for future development are being made. The importance of such projects of self-creation for the narrative constitution of the self has been stressed by Jonathan Glover (1988) and fruitfully applied to bioethical issues by David DeGrazia (2005: Ch. 3).

In conclusion, while narrative approaches to personal identity initially may appear to be rather "liberal" in that they do not come up with necessary and sufficient conditions for persons to be and remain who they are, a closer look reveals that some accounts place substantial constraints on the kinds of self-narration that can be properly regarded as constituting a person's identity.

Normative Implications of Narrative Accounts of Personal Identity

Several proponents of narrative accounts of personal identity put their theories to test by assessing the ethical concerns raised with regard to the psychological effects of brain interventions in general and DBS in particular. As the earlier discussion of Schechtman's evaluation of Mr. Garrison's fictive case already implied, she considers DBS to pose a general threat to personal identity by causing "a disruption of the narrative flow of a life" (2010, p. 137). The reason why DBS appears as a mitigating factor to Schechtman in deciding about the responsibility of Mr. Garrison for commitments of his earlier "unstimulated" self is that the rapidity with which DBS "mechanically" induces psychological changes runs afoul of the articulation constraint. While Mr. Garrison may try to establish narrative continuity with his self before DBS by offering reasons for his changed ways of thinking and behaving, doubts may always lurk in the background as to whether they are the real reasons for his psychological changes or whether, in fact, they have been caused by brain stimulation. In a helpful manner, Schechtman's narrative self-constitution view not only offers a reason for ethical scrutiny with regard to DBS, but also hints at possibilities to avoid the threat that DBS may pose to personal identity: "Since narrative is a dynamic notion, continuity of narrative is thoroughly compatible with even quite radical change. The important thing is that the change be understood in a way that makes it part of a coherent personal narrative, one that patients and their close associates can see as, overall, self-expressive and self-directed"

(Schechtman 2010, p. 138). To the extent that the associates of a patient are able to offer him interpretations of his psychological transformation that make sense in the overarching narrative framework of his life, they can help to avoid damage to his identity.

Baylis, whose relational account of personal identity has many commonalities with Schechtman's narrative theory, also acknowledges that "DBS can be a uniquely disruptive experience resulting in dramatic changes in behavior, mood and cognition" and that these changes can have a major impact on personality (Baylis: 2011). Baylis insists, however, that changes in personality are not in themselves sufficient to warrant claims about threats to personal identity. Baylis emphasizes the dynamic nature of narrative concepts of the self and argues that if DBS is to be seen as a threat to the process of identity formation simply because it results in significant personality changes, then it would be legitimate to characterize all significant life events (both positive and negative events) such as graduation, job loss, marriage, birth of a child, divorce, and earthquake as threats to personal identity insofar as they could result in dramatic changes in personality (Baylis 2011). According to Baylis, it is only "when DBS undermines agency to such an extent that the person is no longer able to meaningfully contribute to the authoring of her own life," that it may be accurate to describe DBS as a threat to personal identity.

Our discussion of the concept of forensic personhood has shown that cases in which brain interventions lead to a loss of impulse control or other prerequisites of autonomous agency do not require an assessment in narrative terms. It may provide stronger reasons for ethical concern to describe such cases as involving a threat to the patient's continued existence as a (forensic) person, rather than saying that his narrative identity is being threatened. However, the benefit of viewing the consequences of brain interventions through a narrative lens may become more obvious if the psychological changes experienced by patients are more ambiguous than in the extreme cases in which personhood seems to be undermined. Narrative approaches to personal identity prove particularly useful in cases like our second case report which seem to involve personality changes on the part of the patient.

Personality changes pose a particular challenge for any ethical assessment of brain interventions as they are complex and hard to measure and at the same time characterized by a high life impact (cf. Müller and Christen 2011, p. 8). On a narrative account of personal identity, personality changes do not appear inherently problematic because any meaningfully unfolding life story includes profound psychological changes in the wake of successful and sometimes failing projects of self-creation (cf. DeGrazia 2005, p. 236). This seems to justify the following conclusion by Matthis Synofzik and Thomas Schlaepfer: "Thus, the ethically decisive question is not whether DBS alters personality or not, but whether it does so in *a good or bad way* from the patient's very own perspective" (2008, p. 1514; emphasis in original). Noteworthy, Synofzik and Schlaepfer arrive at this conclusion using different theoretical premises. Rather than subscribing to a narrative concept of personal identity, they rely on an understanding of personality "as a supramodal representational system with largely heterogeneous

functional and (self-)representational levels" (2008). Whatever the theoretical merits of such a "naturalistic account of the self" may be, it certainly does not offer any grounds for the ethical evaluation of personality changes (cf. Witt et al. 2011). All that may be derived from it is a principle of caution, as each and every neurotechnical intervention *can* have an impact on personality, even if it primarily aims at altering sensorimotor circuits as in DBS for PD, because among the important parts of personality are also "low-level sensory, motor or vegetative states" (ibid.).

In contrast with this plausible but rather unspecific result, thinking about personality changes in narrative terms opens up a new evaluative perspective in that not all features of a person's personality are of equal importance to her in view of her narratively constituted self-concept. For instance, it may seem less problematic to potentially alter a particular personality trait in a person for whom this characteristic plays a minor role in his self-narrative, than in someone for whom it represents a core feature of his self-concept. While the protection of such core features of personality should have high priority in risk-benefit considerations concerning brain interventions, no personality trait has to be declared sacrosanct on narrative grounds.

History of humankind is ripe with examples of people sustaining the narrative thread of their selves in spite of the most far-reaching disruptions in their life stories. Narrative accounts can provide an explanation for the individual differences in the ability to accommodate disruptive events in a stable self-concept by pointing out that self-constituting narratives may differ in their pliability because of their specific contents and because of the varying rigidity of their *narrative styles* (cf. Merkel et al. 2007, p. 260). Moreover, as the discussion of Schechtman (2010) has shown, narrative approaches may even indicate new ways of handling cases of "identity crisis" after brain interventions in that it may be helpful to mobilize "narrative resources" for their integration into a meaningful life story not only in the patient himself, but also in his significant others.

Finally, although this possibility has not been explored so far, it may be worth considering the implications which a narrative account of personal identity may have for patient information on possible psychological side effects in the consent process for brain interventions. It may turn out to be a real challenge to convey to a patient the subtle ways in which his self-concept may change in the wake of a brain intervention including, for instance, changes in his value system to the effect that even his current standards of weighing the possible consequences of treatment may change. At any rate, the physicians involved in clinical decision making for brain interventions should try to take the individual differences in the way persons conceive of themselves and of life itself into account. The currently available instruments for the psychological assessment of personality changes and the psychiatric assessment of personality disorders do not seem to pay due attention to possible changes in a person's self-concept which, according to Merkel and coauthors' "narrative account of personality" (2007: 5.4.4), represents an integral part of personality. The infinite diversity of self-narratives calls for an individualized assessment of the psychological sequelae of brain interventions.

It is thus no accident that Schüpbach and colleagues (2006), who created much interest by describing the unexpected dissatisfaction in PD patients who gained considerable relief from their motor symptoms by DBS, were among the first to employ unstructured in-depth interviews in their assessment of the psychosocial consequences of DBS. A narrative approach to personal identity thus requires that the validated quantitative test instruments for the detection of personality changes should be complemented by qualitative methods like semi-structured open interviews. These interviews should not only be conducted with the patients, but also with their relatives and caregivers, because only by taking different perspectives into account can points of disagreement be identified between a patient's self-concept and the way others conceive of her. Once again, this conclusion receives independent support from authors who themselves do not understand identity in narrative terms (cf. Synofzik and Schlaepfer 2008, p. 1515; Witt et al. 2011).

Conclusion

After brain interventions people may experience various kinds of psychological changes, some of which are more relevant for the ethical assessment of such interventions than others. According to many neuroethicists, a particularly important class of psychological effects is provided by changes that affect the identity of persons. Discussions of the possible impact of brain interventions on personal identity usually assume that such an impact is cause for concern, not hope. However, this evaluative assumption holds true only for some of the accounts of personal identity reviewed in this chapter. It seems justified as far as the numerical identity of persons is concerned. Something undoubtedly went wrong, if after a brain intervention a patient either lost one of the capacities necessary for personhood or can no longer be recognized as the same person in the numerical sense. The possibility of a person's existence coming to an end with the corresponding human being staying alive cannot be accommodated by all accounts of personal identity. In particular, those accounts that adopt biological criteria for the persistence of persons leave no room for such issues of numerical identity. The concept of forensic personhood, on the other hand, not only allows for human beings who (at different stages of their lives) cannot be recognized as forensic persons, but also makes sense of the possibility of personal identity change where two persons (or possibly even more persons as in cases of dissociative identity disorder, cf. the ► Chap. 24, "Dissociative Identity Disorder and Narrative") consecutively "reside" in the same body. As both possibilities entail the (transient or permanent) loss of something valuable, i.e., the capacity for autonomous agency, the framework of forensic personhood justifies the evaluative assumption that any impact on numerical personal identity has to be considered as a risk and never as a benefit of brain interventions.

If, in contrast, personal identity is construed narratively, then it seems implausible to assume that any change in a person's narrative identity must necessarily be for the worse. The concept of narrative identity is closely related to the concept of

personality in that both concepts can be readily employed to answer the characterization question related to persons. It can thus be understood why narrative approaches to personal identity hold a lot of promise for the ethical evaluation of personality changes occurring after brain interventions. In some sense, personality changes definitely affect who a person essentially is, which is why their assessment is so important in clinical decision making. On most accounts of personality, however, persons and their significant others may be afraid of some personality changes and welcome others. While there may be good reason to generally ban certain modifications of personality, e.g., enhancing aggressiveness, it does not seem in principle objectionable to modify personality by brain interventions. Narrative approaches to personal identity may offer guidance for risk-benefit considerations related to personality changes. Such treatment decisions are particularly challenging, irrespective of whether personality changes are envisaged as possible side effects or as the intended outcome of an intervention.

Ultimately, both approaches to personal identity prove useful for the ethical assessment of brain interventions. On the one hand, considering the possibilities of losing personhood and of (numerical) personal identity change helps identifying illegitimate goals and grave risks of brain interventions. Thinking about personal identity in narrative terms, on the other hand, suggests an ethically significant sense in which a person may change without her existence thereby coming to an end and hints at useful criteria for the assessment of such changes.

Cross-References

▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Dissociative Identity Disorder and Narrative
▶ Ethical Implications of Brain Stimulation
▶ Ethical Implications of Cell and Gene Therapy
▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
▶ Ethics in Neurosurgery

References

Baylis, F. (2011). "I am who I am": On the perceived threats to personal identity from deep brain stimulation. *Neuroethics*, *4*. doi:10.1007/s12152-011-9137-1

Boer, G. J. (1994). Ethical guidelines for the use of human embryonic or fetal tissue for experimental and clinical neurotransplantation and research. *Journal of Neurology, 242*, 1–13.

DeGrazia, D. (2005). *Human identity and bioethics*. New York: Cambridge University Press.

Focquaert, F. (2009). Direct intervention in the brain: Ethical issues concerning personal identity. *Journal of Ethics in Mental Health, 4*(2), 1–7.

Glannon, W. (2009). Stimulating brains, altering minds. *Journal of Medical Ethics, 35*, 289–292.

Glannon, W. (2010). Consent to deep-brain stimulation for neurological and psychiatric disorders. *The Journal of Clinical Ethics, 21*, 105–112.

Glover, J. (1988). *I: The philosophy and psychology of personal identity*. London: Penguin.

Leentjens, A. F. G., Visser-Vandewalle, V., Temel, Y., & Verhey, F. R. J. (2004). Manipuleerbare wilsbekwaamheid: een ethisch probleem bij elektrostimulatie van de nucleus subthalamicus voor ernstige ziekte van Parkinson [Manipulation of mental competence: an ethical problem in case of electrical stimulation of the subthalamic nucleus for severe Parkinson's disease]. *Nederlands Tijdschrift voor Geneeskunde, 148*, 1394–1398.

Lipsman, N., Zener, R., & Bernstein, M. (2009). Personal identity, enhancement and neurosurgery: A qualitative study in applied neuroethics. *Bioethics, 23*(6), 375–383.

Locke, J. (1975). *An essay concerning human understanding*. In P. H. Nidditch (Ed.). Oxford: Clarendon.

Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B., & Rosahl, S. (2007). *Intervening in the brain. Changing psyche and society*. Berlin: Springer.

Müller, S., & Christen, M. (2011). Deep brain stimulation in Parkinsonian patients – Ethical evaluation of cognitive, affective, and behavioral sequelae. *AJOB Neuroscience, 2*(1), 3–13.

Schechtman, M. (1996). *The constitution of selves*. Ithaca: Cornell University Press.

Schechtman, M. (2009). Getting our stories straight: Self-narrative and personal identity. In D. J. H. Mathews, H. Bok, & P. V. Rabins (Eds.), *Personal identity and fractured selves: Perspectives from philosophy, ethics, and neuroscience* (pp. 65–92). Baltimore: Johns Hopkins University Press.

Schechtman, M. (2010). Philosophical reflections on narrative and deep brain stimulation. *The Journal of Clinical Ethics, 21*(2), 133–139.

Schechtman, M. (2011). The narrative self. In S. Gallagher (Ed.), *The Oxford Handbook of the self* (pp. 394–416). Oxford: Oxford University Press.

Schüpbach, M., Gargiulo, M., Welter, M. L., Mallet, L., Béhar, C., Houeto, J. L., Maltête, D., Mesnage, V., & Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*(12), 1811–1816.

Synofzik, M., & Schlaepfer, T. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*, 1511–1520.

Witt, K., Kuhn, J., Timmermann, L., Zurowski, M., & Woopen, C. (2011). Deep brain stimulation and the search for identity. *Neuroethics*, *4*. doi:10.1007/s12152-011-9100-1

Extended Mind and Identity

26

Robert A. Wilson and Bartlomiej A. Lenart

Contents

Abstract

Dominant views of personal identity in philosophy take some kind of psychological continuity or connectedness over time to be criterial for the identity of a person over time. Such views assign psychological states, particularly those necessary for narrative or autobiographical memory of some kind, and special importance in thinking about the nature of persons. The extended mind thesis, which has generated much recent discussion in the philosophy of mind and cognitive science, holds that a person's psychological states can physically extend beyond that person's body. Since "person" is a term of both metaphysical and moral significance, and discussions of both extended minds and personal identity have often focused on memory, this article explores the relevance of extended cognition for the identity of persons with special attention to neuroethics and memory.

R.A. Wilson (✉) • B.A. Lenart
Department of Philosophy, University of Alberta, Edmonton, AB, Canada
e-mail: rwilson.robert@gmail.com; b.a.lenart@gmail.com; lenart@ualberta.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_14,

Personal Identity

Philosophical work on personal identity has involved answering two related questions: what is the nature of persons, and what criteria identify a person over time? Answering this second question, the question of diachronic identity, presupposes an answer to the first question. We need to know what persons are in order to identify them over time.

Standard views of personal identity have tended to assume, following John Locke's watershed work on personal identity in *An Essay Concerning Human Understanding*, that defining the nature of persons (or "personhood," as philosophers are apt to say) requires understanding the nature of the self, where the self is constituted by a unified continuity of conscious states. This assumption about the nature of persons underpins contemporary psychological accounts of a person's identity over time, most of which draw on Locke's own appeal to memory. The memory one has of one's self is one's narrative or autobiographical memory. Such accounts, sometimes referred to as "neo-Lockean," have held sway as the dominant view of diachronic personal identity throughout the twentieth century.

Locke himself proposed memory as criterial for tracking a person through time because he viewed "person" as:

> a forensic term appropriating actions and merit; and so [belonging] only to intelligent agents capable of a law, and happiness and misery. This personality extends itself beyond present existence to what is past, only by consciousness; whereby it becomes concerned and accountable, owns and imputes to itself past action, just upon the same ground and for the same reason that it does the present. (1690, Bk. II, Ch. XXVII, §26)

That is, for Locke, the term "person" functions to allow us to praise, to assign blame, and to hold individuals accountable for what they have done in the past.

As the passage makes clear, for Locke the designation of personhood is to be attached to a self. Given Locke's earlier discussion of the identity of the same substance, same human being, and same person over time (II:xxvii.6), it is clear that Locke takes this self to be something other than a substance or a kind of living organism. For Locke, the self is distinctively individuated over time by some kind of psychological continuity, one in which memory plays a key role.

One influential argument that Locke provides for this view turns on his introduction of a now-famous example. Imagine a prince and a cobbler, with their very different personalities, characteristics, and memory, waking up one day with the minds of each of them switched into the body of the other (II:xxvii.15). If these mind-body pairs were switched, Locke explains, then the personality and memories of the prince would be transferred into the cobbler's body. This would be a case, Locke argues, in which the prince now exists in the cobbler's body, and the cobbler in the body of the prince. The reason for this is that memory and the cognitive capacities required for memory serve the forensic function Locke attributes to personhood. Locke writes "whatever past actions it [the person] cannot reconcile or appropriate to that present self by consciousness, it can be no more concerned in than if they had never been done" (II:xxvii.26).

Despite the dominance of Lockean and neo-Lockean views of persons and their identity over time, they are not the only game in town. Theories proposing a biological criterion for diachronic identity discount the importance of psychological continuity, looking instead to the structural integrity of an organism as the basis for an individual's identity through time. Proponents of the biological criterion take an essentialist approach to the problem of identity by pursuing the question of what kind of thing humans are, basing their account of personal identity on the persistence conditions of that thing. For instance, Eric Olson (1997, 2003) argues that human beings are essentially biological organisms or human animals. Thus for Olson, tracking the identity of a person has nothing particularly to do with that person's mental states or memory. Instead, it is sufficient to know the persistence conditions of human organisms and to apply them to a person of interest in order to identify or reidentify that person over time.

Morally speaking, we ought to feel apprehensive about a view that would be willing to hold a person accountable even if its mind were unaware of the body's previous activities and, in fact, cognizant of entirely different actions. The full weight of this implication is apparent if we imagine an assailant and his or her victim switching brains. According to the biological approach, the assailant's body should be punished for the crime despite the fact that the experience of the punishment is to be added to the victim's already emotionally traumatized conscious states; the victim would remember being assaulted and then would be punished for his or her assailant's actions. Thus, if we are to preserve both the connection between person and moral blame that Locke made in viewing "person" forensically and our intuitions about this particular case, we seem drawn from the biological to the psychological criterion.

Another essentialist alternative to neo-Lockean views postulates a continuing mental substance, the identity of which may be independent of either psychology or biology. Although some version of a substantive view of personal identity may be widely accepted by the general public, such views are at best marginally represented among philosophers. The problem making this approach philosophically unpalatable is this: if the substance is not altogether lacking in psychological and biological traits, then its identification will inevitably be based on those traits, which will in effect amount to a trimmed (and likely somewhat impoverished) psychological or biological view. Thus, such a view, in order to represent a distinctive alternative, would have to postulate a featureless substance. However, as David Shoemaker has argued, "[w]e cannot track immaterial egos floating free from any particular psychological properties, so on this view we would never be justified in claiming to have re-identified anyone" (2012, Sect. 2.5).

Despite the departure of neo-Lockean views from a substance-based account of persons, their focus on consciousness and memory shares with such accounts an attempt to differentiate humans from nonhuman animals and plants by appealing to the nature of their mind or soul. Aristotle famously argued that although humans share in the nutritive and perceptive capacities of plants and nonhuman animals, the rational soul is unique to human beings. Although Locke eschews a commitment to any kind of substance as the grounds for his account of personal identity, Locke's views do

reflect the ratio-centrism that one finds in this Aristotelian tradition, as well as a certain kind of individualism about the nature of persons and their identity over time. This is because certain rational cognitive capacities are required for the autonoetic formation of episodic memories that is tied to orthodox conceptions of personal identity and personhood, and those capacities are conceptualized as depending solely on aspects of the individual herself. These forms of ratio-centrism and individualism remain features of neo-Lockean views of persons in the contemporary literature.

There are morally troubling results of both the ratio-centrism and individualism of neo-Lockean views of personal identity, especially in combination. First, such views seem to imply that individuals with certain cognitive limitations (such as, for example, a relatively limited ability to track their own personhood through time) cannot claim the right-conferring status of personhood. This resultant depersonalization of the "mentally deficient" amounts to their subhumanization and with it an abandonment of the universal ascription of fundamental human rights. Second, as a consequence of the inherent individualism of neo-Lockean views, external resources – such as other people, environments, or technologies – that may be intrinsic to certain cognitive (and other) capacities, are viewed as extrinsic to an individual's status as a person. Such external resources may causally enhance the cognitive capacities of an individual who would otherwise fall below the cognitive threshold for full personhood set by neo-Lockean views. Despite this, such external resources cannot themselves be the basis for the moral status of persons.

While the links between narrative memory, rationality, and the self make the ratio-centric and individualistic biases in psychological accounts of identity of persons over time and thus the nature of persons readily identifiable, such biases are also implicit in biological accounts of persons. Faced with the problem of specifying which of the many kinds of unified, living, persisting organisms are persons, proponents of biological views also tend to fall back on appeals to the kinds of consciousness and rationality possessed by typically human organisms. Thus, when lines must be drawn, rationality and one's intrinsic capacities become the historically authoritative boundary markers for personhood.

The Empirical Study of Mind and Personal Identity

For the most part, traditional work on personal identity in philosophy has proceeded with little reference to, let alone sustained discussion of, empirical work in the clinical and cognitive sciences. There are at least four areas, however, in which such work has been seen to be relevant to ongoing philosophical discussions of personal identity, particularly by those influenced by neo-Lockean, memory-based criteria for personal identity. These focus on clinical phenomena that in certain respects parallel fictional and philosophical fantasies, such as Dr. Jekyll and Mr. Hyde, or Locke's own "day man" and "night man," in raising questions about the relationship between persons, minds, and bodies, as well as about the identity of persons over time. Similar questions are nascent in more recent discussions of enhancement technologies.

Split-Brain Cases. These are chiefly examples in which patients have undergone the surgical procedure of commissurotomy, which severs the corpus callosum that provides the primary neurological channel connecting the left and right hemispheres of the brain. Commissurotomy as a surgical technique was first performed on patients with severe forms of epilepsy featuring grand mal seizures in 1940. Split-brain cases were initially described by the Nobel Prize winning neuropsychologist Roger Sperry (1966, 1968a, b), following work that he had undertaken with the neurosurgeon James Bogen at the California Institute of Technology. Sperry's characterizations provoked much discussion in the philosophical literature on personal identity in the 1970s and 1980s (see Marks 1981; Dass 1995).

The initially surprising finding was that patients who had undergone commissurotomy manifested recognitional behaviors that differed markedly, depending on how those behaviors were elicited. For example, when presented with a visual stimulus shown tachiscopically for 150–200 ms in the left of the visual field, and asked both if they had seen anything and to describe what they had seen, patients showed no awareness of these stimuli; when probed to draw or guess at what was presented, however, such patients performed significantly better than chance. The explanation for this perhaps innocuous-sounding discrepancy is that the left visual field, which is processed in the right hemisphere, has a functional specialization for imagery and nonverbal processing. Since the primary signaling channel between the hemispheres is absent following commissurotomy, there is information available to the right hemisphere that is isolated from the left hemisphere, which has a functional specification for language and categorization. Hence, a probe that draws on the left hemisphere, which lacks information about the stimulus, will elicit non-recognitional behaviors; a probe that draws on the right hemisphere, which does possess that information, elicits recognitional behaviors. Further discrepancies found between how patients reported information acquired through tactile and visual modalities could be explained in a similar manner.

Sperry claimed that results such as these suggested that "the surgery [commissurotomy] has left these people with two separate minds, that is, two separate spheres of consciousness" (1966, p. 299). Reports of these findings of split-brain cases were thus sometimes interpreted as cases in which there were two persons in one body, with the Canadian philosopher Roland Puccetti arguing further that this is the proper way to think about persons and their bodies more generally (Puccetti 1973a, b, 1981).

Psychiatric Disorders Involving the Disorder of the Self. Perhaps the most prominent psychiatric disorders involving the self are dissociative disorders, including multiple personality disorder, which was recognized in the Diagnostic and Statistical Manual of Mental Disorders (DSM) in both its second and third editions, published, respectively, in 1968 and 1980. This medical conception of the nature of the disorder, which built on the popular idea that a given human body may well possess more than one personality and that these personalities can govern the behavior of that body in very different ways, comported with the kind of speculation fueled by the work on "split-brain patients" in that it seemed also to lend itself to a philosophical gloss of there being at least "two persons in one body," each causally responsible for directing the behavior of that body at different times (cf. Braude 1991). The substantial

reconceptualization of MPD as "dissociative identity disorder" in DSM-IV (1994) as requiring "the presence of two or more distinct identities or personality states" that alternately control the individual's behavior, and that manifest relatively cohesive narrative memories that are isolated from one another, in effect suggests that disintegration of the self, rather than its multiplication, is at the heart of the condition. Perhaps "multiples" have less than, rather than more than, one self (see Hacking 1995).

Memory Loss Over Time. The third cluster of empirical phenomena that philosophers thinking about personal identity in the neo-Lockean tradition have appealed to concern cases of extreme, even if gradual, memory loss over time, such as one finds amongst Alzheimer's patients and others suffering from age-related forms of dementia. Central to such afflictions is the loss or severe diminution of memory, not simply narrative memory of one's past but of the ability to recognize one's family or close friends and one's even quite recent interactions with them, as well as the abilities to remember and act on one's own plans and expressed desires, and procedural memory for knowing how to perform actions, such as driving a car or riding a bicycle. As such abilities decline, so too does one's capacity for a cohesive mental life, calling in to question the relationship between one's self at distinct times, such as the past and present or the present and future (de Grazia 2005, Chap. 5).

Enhancement Technologies. More recently, some discussions of personal identity have shown sensitivity to developments in the clinical sciences concerned with cognition and the mind that focus not so much on traditional pathologies but on enhancement technologies. For example, Carl Elliott (2003) has discussed the variety of ways in which technologies – ranging from accent reduction training and other forms of voice modification through to the psychopharmacological mood adjustments induced by drugs such as Prozac that lead some users to describe themselves as feeling "better than well" or as finally being able to "be themselves" – have been developed and used to influence one's sense of narrative identity over time. Some such technologies, such as cochlear implants and prosthetic limbs, literally augment the brain and body of a person in ways that either restore missing or lost capacities or enhance such capacities beyond those possessed by the fictional normal persons. Although these discussions have been typically cast in terms of the cultural and scientific significance that such technologies have for conceptions of the self, public policy, and individual lifestyle, they remain relatively undigested in the literature on personal identity. Likewise, consider the more science fictional projections of transhumanists who are focused on the possibility of substantial life span extensions that involve technologies allowing for the downloading of human minds into new bodies or even nonbiological forms of instantiation (Kurzweil 2005; Agar 2010). Both the presumptions of and implications for such possibilities vis-à-vis personal identity have received some recent discussion (Schneider 2009).

Extended Cognition and Extended Minds

The empirical work drawn on in discussions of personal identity recounted above has tended to reflect the predominance of psychological and more particularly

memory-based views of personal identity. Such views have taken the brain and neural activity to be distinctive, vis-à-vis personal identity, from merely bodily activity. In this section we turn to views of cognition, including of memory, that question whether neural activity itself is sufficient for cognitive processing of particular kinds or even for having a mind. According to proponents of the extended mind thesis (Clark and Chalmers 1998), or the hypothesis of extended cognition (sensu Rupert 2009), the answer to these questions is "no."

The extended mind thesis holds, perhaps counterintuitively, that cognition does not take place exclusively in the head of the cognizer. As such, it is a form of externalism about the mind or cognition that developed in the 1990s as part of a longer dialectic between individualists (Fodor 1987; Segal 1989; Egan 1991) and their externalist critics (Burge 1979; Shapiro 1993; Wilson 1994). In contrast to the debate between individualists and externalist to that point, which had focused on the notion of mental content or representation, proponents of the extended mind argued that minds or cognitive systems themselves were not fully located within the bodily envelope of the individual (Wilson 1995; Clark and Chalmers 1998). On this view, features of, or structures in, an organism's environment could in principle be, and sometimes in practice were, physical constituents of that organism's cognitive systems. Such cognitive systems are extended in that they do not begin and end at the skull or even body of the individual cognizer. The extended mind thesis can be readily motivated theoretically, as well as by reflection on everyday ways in which we rely on and even come to incorporate parts of our artifactual environment into our cognitive activities.

Theoretically, the possibility of extended cognition follows from functionalism in the philosophy of mind, where what matters for cognition is not the what or the where but the how of cognition. For at least a sophisticated form of functionalism, cognitive processing is, in essence, a matter of a certain kind of structural and dynamic functional organization. Given the commitment to materialism shared by most functionalists, functional organization is physically realized and so physically located. But what does the realizing, and just where that matter is located, is of secondary importance. Networks of neurons organized in certain ways can realize particular cognitive systems, but so too might silicon chips so organized. And the physical stuff realizing such networks is often located inside a skull, but it may also be distributed between head and world. Thus, certain kinds of parity considerations lie at the theoretical heart of the idea of extended cognition, ones that appeal to functionalist commitments that we view as running deep in the cognitive sciences (Clark and Chalmers 1998; Wheeler 2010; Wilson in press).

In terms of reflection on everyday cognitive activities, consider our systematic reliance on pen-and-paper calculation in order to solve even minimally complicated multiplication problems. Here we store intermediate solutions on the paper, using perception and action to mediate information flow between the symbols stored in our heads and those stored on the paper. The cognitive process of solving a multiplication problem, in this case, involves integrated information processing both inside and outside of the person's body. Moreover, as the workload involved in many cognitive tasks increases – more information to store and track, higher

attentional demands, more levels to decision-making – the corresponding cognitive processing comes to systematically depend on the smooth integration of in-the-head cognition with cognitive tools and structures outside of the head. Proponents of extended cognition take such examples of cognitive offloading to point to how extended cognitive systems have been shaped evolutionarily, developmentally, and culturally for everyday cognitive tasks (Clark 2008; Wheeler 2005; Wilson and Clark 2009). Cognitive scientists adopting a "situated cognition" perspective on a variety of topics, such as problem solving (Kirsh 2009), learning (Sawyer and Greeno 2009), and rational decision-making (Brighton and Todd 2009), continue to explore systematically the role that such offloading plays in everyday cognition.

A more precise statement about extended cognition that concerns particular cognitive activities and that makes explicit the idea of an extended cognitive system summarizes this overview of the extended mind thesis:

> A cognitive activity is extended just if it is generated or sustained by the activity of one or more extended cognitive systems.
> A cognitive system is extended just if it contains, as physical constituents, one or more processing resources that are not contained inside the head or body of that individual.

Multiplication performed by a person using pen and paper involves extended cognition, provided that (a) the pen and paper function as processing resources that are (b) not contained inside the head or body of that person, and (c) are physical constituents of a cognitive system which (d) generates or sustains that activity. Those resistant to the idea of extended cognition and the extended mind can be viewed as rejecting one or more of these provisos, most commonly (a) or (c) (Adams and Aizawa 2008; Rupert 2009).

Although the extended mind thesis was originally articulated as a merely possible alternative to the view that cognition takes place entirely in the head, the thesis has come increasingly to be defended as a plausible view of much actual cognition. As such, it has appealed to ongoing work in the cognitive sciences in support of this claim (Wilson 2004; Clark 2008; Wilson and Clark 2009; Wilson in press), including the use of gesture for linguistic communication (Clark 2008), action-guided views of perception (Wilson 2010), and memory (Wilson 2005). Given that memory has played a prominent role in discussions of both the extended mind and personal identity, we make that our focus below.

External memory storage for problem solving, planning, and decision-making features in both the multiplication example we have discussed as well as Clark and Chalmer's (1998) classic Otto-Inga thought experiment in which one person, Otto, compensatingly comes to rely on and utilize a notepad as effectively as another person (Inga) uses internal memory storage for finding one's way to a particular location in a city. The kind of parity considerations in play here can be used to motivate a broader rethinking of the kind of memory central to personal identity.

While there are many ways in which memory has been conceptualized – short term vs long term (Atkinson and Shiffrin 1968), episodic vs semantic (Tulving 1972, 1983, 2010), procedural vs declarative (Graf and Schacter 1985;

Schacter and Tulving 1994; Schacter 2010), and iconic vs linguistic (Sperling 1960; Neisser 1967) – as we have seen, it is narrative or autobiographical memory that is most directly relevant to discussions of personal identity. The sense of having a continued psychological existence over time, such that one can remember oneself having done certain things in the past, matters to us and is what allows us to guide our current actions and plan our futures in light of who we are.

Such narrative or autobiographical memory, particularly in its individualistic guise, might be thought to fall under the broad umbrella of declarative (vs procedural) memory since it is, more specifically, a type of episodic memory involving autonoetic (or self-knowing) awareness (Tulving 2010). Autonoetic recollection of an event is essentially a reexperiencing of a past experience, making episodic memory a clear example of tracking a self – one's own self – through time and thus well suited for tracking personhood over time.

Given the extended mind thesis, however, narrative memory need not be bound exclusively to individualistic recollection but can come to incorporate the world beyond the individual in a variety of ways. One such way involves cognitive offloading. In fact, Daniel Dennett's "Making Things to Think With" illustrates our habit of offloading cognitive tasks into the environment with the example of elderly people who are incapable of recalling simple daily routines and suffer from other memory-related deficiencies once they are housed in institutions such as nursing homes (1996, pp. 134–139). Many such signs of dementia are less pronounced or disappear altogether once people are returned to their own homes where they have offloaded many of their daily routine schedules (such as taking their medications) on items or places that remind them of what they have to do, how they ought to do it, and other kinds of pertinent information.

Even though narrative memory has typically been conceptualized individualistically, as with other forms of extended memory, it can come to integratively rely on aspects of familiar environments, as in Dennett's example of cognitive offloading (cf. also Wilson 2004, pp. 189–198). But extended narrative memory also departs from individualism in another way: it can be shared and co-constructed by two or more individuals (Wilson 2004, pp. 191, 207–211; cf. Barnier et al. 2008).

This second dimension to extended narrative memory might be thought to call into question a putatively clear-cut distinction between autobiographical and collective memory. Collective memory has received much discussion over the past decade or so in the humanities and social sciences, especially in Holocaust and trauma studies (see, e.g., Olick, 2011; see also Wilson 2005, Theiner 2008, 2013). Collective memory is often commemorative of significant past events, ritualistic, and political in nature. For example, we collectively remember atrocities on the calendar date on which they were committed or engage in joint actions that express our political affinity and sympathy with (other) victims of a crime or natural disaster.

We think that we can maintain a version of the distinction between narrative and collective memory by thinking of the sharing of one's narrative memories in the same way that we can think about the offloading of those memories. Integrating things in one's immediate environment to form an extended memory system is a

form of extended cognition utilized by people with Alzheimer's disease and other neurodegenerative conditions affecting memory. Here it is the individual who remembers, but the activity of remembering is extended, being distributed between that individual and things in her environment. Likewise, when one's narrative memories involve a co-participant, it is still one's self who remembers, even if the activity of remembering is socially extended, being distributed between the individual and her co-rememberer. What one needs, in both cases, is some kind of asymmetry between the person whose autobiography is being actively constructed and the things or other persons involved in that construction (for one account, in terms of the notion of locus of control, see Wilson 2004, pp. 184–187, 197–198).

In contrast to such cases of extended narrative memory, in cases of collective memory it is some kind of collective or group that remembers, distributing the task of remembering between different individuals within a group in ways that make it implausible to identify any one of them as "the" person who remembers. In fact, what is remembered in collective memory is not autobiographical, even if it involves things that have happened to particular individuals (or even just one individual). It is remembering that is (typically) episodic but not autobiographical.

The relationships between extended and collective memory require exploration beyond our necessarily brief comments here. One general claim that has been made about "group minds" that may prove relevant here is that many putative examples of collective cognition are more plausibly viewed as cases in which the extended cognition of the individual involves a social environment involving other people. This social manifestation thesis – "the idea that individuals engage in some forms of cognition only insofar as they constitute part of a social group" (Wilson 2005, p. 229) – can be applied to memory and viewed as offering both a challenge to proponents of group minds and potentially, at least, an expanded role for the extended mind thesis (see also Wilson 2004, Chaps. 11–12; Barnier et al. 2008). That expanded role contains implications for personal identity.

Extended Personal Identity

An externalist neo-Lockean account of identity is not as puzzling as it may initially sound, especially given that the psychological account appeals so directly to narrative memory. As Alasdair MacIntyre (1984) has observed, people are essentially storytelling animals. The narrative tools we employ to make sense of our identities arise in cultural, historical, and institutional settings. When we take memory seriously in the context of personal identity, it becomes clear that individual identities, just like individual memories, are realized within the context of collective narratives. Individual memories may well serve as the vehicles for individual identities. But such memories are influenced by collective narratives, thus making individual identities heavily reliant on the collective or social contexts within which individuals exist. An appreciation of this relationship between individual rememberers and the collective narratives in which they are immersed

should not only compel us to rethink our understanding of memory, but should also inform our conception of personhood.

The intimate connection between individual and collective remembering has been noted by researchers studying memory since the 1930s. For example, F. C. Bartlett (1932) argued that interests, in the broad sense, taken to mean the development of a person's mental life, are responsible for what a person remembers. Moreover, Bartlett argued that interests themselves have a social origin (p. 256) in customs, institutions, and traditions, which constitute a lasting social schema (p. 264). Rephrased in the language of the social manifestation thesis, Bartlett's argument is as follows: remembering is private and subjective insofar as the individual doing the remembering does so privately. However, all remembering is made possible and is shaped by the social constructions and contexts in which the remembering occurs.

This kind of relationship between individual narratives and collective remembering suggests that the extended mind thesis may be well positioned to augment traditional neo-Lockean views of personal identity. Suppose that the cognitive capacities involved in remembering are not intrinsic to the individual whose identity is being tracked, but are socially manifested capacities. This would imply that a person's identity has a wide realization. Tracking a person's identity over time, on this view, would involve many minds, including the mind of the individual who is tracked. But since "the characterization of wide realizations preserves the idea that properties with such realizations are still properties of individual subjects" (Wilson 2004, p. 141), this externalist view of personal identity does not entail that the individuals who are persons are themselves "wide" or "extended" selves.

Combining a psychological account of personal identity with the extended mind thesis in this way provides one with the resources to solve some of the problems facing standard neo-Lockean views, problems that stem from the individualistic ratio-centrism of such views that we identified earlier. It does so by recognizing narrative-based criteria for personhood that are based on more than just the intrinsic cognitive capacities underpinning the remembering of the normally abled.

The case of patient HM, who suffered from memory loss following bilateral medial temporal lobe resection (Scoville and Milner 1957), serves as an example of an individual whose diachronic identity has been more dramatically socially realized. HM lost the ability to consolidate new information into long-term episodic memories, thereby losing the ability to autonoetically track his own identity over time. An extended account of identity enables a genuine maintenance of personal identity on behalf of individuals who, like HM, are incapable of tracking their own personhood through time.

Recognizing a socially extended realization base for personal identity dovetails with some recent work of Hilde Lindemann on the role that others play in "holding us" – all of us – in our identities. Lindemann (2010) argues that a person's identity is both shaped and preserved by others via the complex interactions between, and varied intertwining narratives remembered and transmitted within, families and other groups. Echoing Dennett on the environmental offloading of cognitive tasks

by proposing that places as well as people can hold individuals in their identities, Lindemann writes:

> It's not just other people who hold us in our identities. Familiar places and things, beloved objects, pets, cherished rituals, one's own bed or favorite shirt, can and do help us to maintain our sense of self. And it is no accident that much of this kind of holding goes on in the place where our families are: at home. (Lindemann 2010, pp. 162–163)

Thus, externalism does not merely change the way we understand the mind; it also affects how we define and track personhood. Selves are a product of both individual and communal processes, and thus personhood should not be defined in solely individualistic terms. To think of personhood as purely individualistic and private is as much a mistake as thinking of memory in such terms.

Maurice Halbwachs once mused that in order to experience private remembering that is minimally influenced by social contexts, we should look to our dreams, which "are composed of fragments of memory too mutilated and mixed up with others to allow us to recognize them" (1941, p. 41). Persons, whose identity is strongly tied to memory, must have widely realized identities that emerge in a social context. They are formed by, held in, and tracked via the memories of others, which themselves are shaped by the collective memories of the various social groups to which these individuals belong.

Concluding Thoughts

The extended mind thesis makes the claim that minds extend beyond the skull. Analogously, the externalist account of personhood might be taken to make the claim that persons themselves are extended in just this way. Although some proponents of the extended mind thesis may indeed be taken to advocate or welcome such a claim (see Clark 2001, 2003; cf. Clark and Chalmers 1998), we have sketched a somewhat less radical view. On our view, what is extended or widely realized is the identity of persons while persons themselves, as the subjects of identity over time, are not extended or wide. An individual's personal identity is, to be sure, an important property of that individual, and it is not determined solely by properties or capacities intrinsic to the body of that individual. But like other properties that individuals have that require external resources to be realized, this extended property is still a property of a spatiotemporally bounded and located individual.

All persons can and do rely on others to maintain a cohesive narrative identity. Individuals with cognitive limitations that create difficulties for their tracking their own identities, thus magnifying the problems that we all face in preserving a coherent conception of ourselves, may depend on others more deeply to maintain such cohesive narratives. This provides one way in which an externalist view of narrative identity allows individuals who have traditionally been viewed as falling below the status of personhood – namely, those with severe or increasing cognitive

disabilities and limitations – to manifest personhood. And it does so without viewing their status as persons as different in kind from that of others. Like the regularly cognitively endowed, their personal identities are socially manifested properties, albeit ones that are more deeply reliant on their social context.

Consider persons who slip gradually, over time, into ongoing states of dementia. On an externalist account of personhood, such persons need not forfeit their identities as their minds and memories deteriorate. This is because even though they gradually lose their capacity for individual memory, their identities are realized in the collective remembering of others. Because the externalist account of personhood does not share in the ratio-centrism that individualistic variants of the psychological approach manifest, it has a greater potential to recognize full personhood in such cases.

Embracing the idea of extended narrative memory is one way to broaden neo-Lockean views in ways that make them apt to be more inclusive about persons. But the shift from an individualistic to an extended mind view also serves to pry neo-Lockean views from their traditional ratio-centrism in other ways as well. One reason why ratio-centrism is so deeply embedded in the psychological account is that most variants of the neo-Lockean approach focus solely on memory as a criterion for diachronic identity. Persons, and more specifically, personalities, however, are constituted by more than merely episodic memories. Robert Nozick (1981), for example, states that "[f]or a life to have meaning, it must connect with other things. . .or values beyond itself" (Nozick 1981, p. 594). Some examples of such meaningful and valuable external relations are relationships with other people, continuing and advancing a tradition, children and families, etc. What all such externally valuable relations have in common is that they are saturated with emotive states. Emotions not only color memories but make some more significant than others. What we remember about ourselves or others, the very narratives that constitute identity, are shaped and made more or less meaningful and thus significant by our affective states during memory formation and recollection.

Emotions, like memories and identities, should be understood externally. As Sue Campbell has pointed out, affective states must be expressible in order to be individuated (1997, p. 66). Campbell's view that what we feel is largely determined by what we express leads to the worry that some individuals can be quite vulnerable to being controlled by others. Campbell explains: "One of the most obvious ways in which our feelings are controlled through their expression is by the power of interpreters to view the occasions of our lives and respond to our expressive acts" (1997, p. 135). Such control over affective states can be easily carried over to controlling large portions of someone's narrative and thereby shaping and constructing an inauthentic identity (see Levy 2007a, b).

Issues of authenticity also emerge in the context of extended identity. Levy argues that an acceptance of the extended mind thesis voids the distinction between neurological interventions "by way of psychopharmaceuticals, transcranial magnetic stimulation, or direct brain stimulation" (2007a, p. 7) and more traditional methods of altering mental states, such as using psychological practices like talk therapy or even enhancing one's nutrition or education (2007a, p. 9). This dissolves

Carl Elliott's (1998, 2003) worry "that if antidepressant use alters my personality traits, it [my personality] is inauthentic, inasmuch as these personality traits cannot be a genuine reflection of who I am" (Levy 2007a, p. 7) since, according to the extended mind thesis, both internal and external interventions are regular occurrences, which contribute to what we consider to be our authentic selves.

The question of authenticity, however, crops up again in the context of an externalist account of *personal identity*, since other people have the power to interpret narratives that constitute a person's identity because "identity maintenance also involves weeding out the stories that no longer fit and constructing new ones that do" (Lindemann 2010, p. 163). Not all such weeding and rewriting of narratives is going to be authentic since facts can be carelessly, as well as purposefully, misinterpreted by others. The worry Campbell raises regarding the power interpreters have over the narratives they interpret is recognized by Lindemann, who recognizes that it is certainly possible to hold someone's identity wrongly or at least clumsily. Narratives must be truthful if they are to genuinely track someone's identity, meaning that the backward-looking narratives that constitute a person's identity must pick out something about that individual that is saliently true. For example, "[i]f you never went to med school, aren't licensed to practice, and don't see patients, then you aren't a doctor, and neither I, nor your doting mother, nor God himself can hold you in that identity" (Lindemann 2010, p. 164). Since inauthentic narratives fail to track individuals genuinely, in effect they mutilate a person's identity and thereby devalue the personhood of the individual.

The externalist account of personal identity thus reveals a fragile side of personhood that remains hidden in individualistic variants of the neo-Lockean approach. Understanding the sensitivity of narratives to interpretative interventions deepens our understanding of what it means to treat people and their identities authentically. Whereas traditionally, personal identity was almost exclusively tied to individualistic episodic memory, the externalist account of personal identity sees narrative integrity, in both its individual as well as collective manifestations, as an essential constituent of diachronic identity. Consequently, extended personal identity not only restores the narrative identities of individuals with severe cognitive disabilities, but also generates a moral imperative toward truthfulness in treating, transmitting, interpreting, and holding of person-tracking narratives.

Finally, extended personal identity acknowledges a variety of people. That is, to the question "what sorts of people are there?," the extended neo-Lockean view points to many more kinds of people than do more traditional accounts. On the extended account, individuals traditionally denied one of the important perks of personhood, namely, *personal identity*, are recognized as having it, despite having cognitive capacities that depart from those typical or normal for (other) persons. Moreover, the extended account of personal identity morally obliges us to protect the authenticity of personal narratives not merely via acknowledgment, but actively (via actual conduct) since we are directly and genuinely responsible for them.

Cross-References

- ▶ Dissociative Identity Disorder and Narrative
- ▶ Ethics of Pharmacological Mood Enhancement
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Neuroenhancement
- ▶ Neuroethics and Identity
- ▶ Neurotechnologies, Personal Identity, and the Ethics of Authenticity

References

Adams, F., & Aizawa, K. (2008). *The bounds of cognition*. New York: Blackwell.

Agar, N. (2010). *Humanity's end: Why we should reject radical enhancement*. Cambridge: MIT Press.

American Psychiatric Association Taskforce on Nomenclature and Statistics. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association Task Force on Nomenclature and Statistics. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association Taskforce on Nomenclature and Statistics. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York: Academic.

Barnier, A., Sutton, J., Harris, C., & Wilson, R. A. (2008). A conceptual and empirical framework for the social distribution of cognition: The case of memory. *Cognitive Systems Research, 9*(1–2), 33–51.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. London: Cambridge University Press.

Braude, S. E. (1991). *Multiple personality and the philosophy of mind*. New York: Routledge.

Brighton, H., & Todd, P. M. (2009). Ecologically rational decision making with simple heuristics. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 264–306). New York: Cambridge University Press.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy, 4*(1), 73–122.

Campbell, S. (1997). *Interpreting the personal: Expression and the formation of feelings*. Ithaca: Cornell University Press.

Clark, A. (2008). *Supersizing the mind*. New York: Cambridge University Press.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, technologies, and the future of human intelligence*. New York: Oxford University Press.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind and Language, 16*, 121–145.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.

Dass, S. (1995). *Commissurotomy and personal identity*. MA thesis, Philosophy, Queen's University, Kingston, Ontario, Canada.

DeGrazia, D. (2005). *Human identity and bioethics*. New York: Cambridge University Press.

Dennett, D. (1996). *Making things to think with* (pp. 135–147). New York: Basic Books. Ch. 4 Kinds of minds: Toward an understanding of consciousness.

Egan, F. (1991). Must psychology be individualistic? *Philosophical Review, 100*, 179–203.

Elliot, C. (2003). *Better than well: American medicine meets the American dream*. New York: Norton.
Elliott, C. (1998). "The tyranny of happiness: Ethics and cosmetic psychopharmacology". In E. Parens (Ed.), Enhancing human traits: Ethical and social implications (pp. 177–188). Washington, DC: Georgetown University Press.
Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge: MIT Press.
Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 11*, 501–518.
Hacking, I. (1995). *Rewriting the soul: Multiple personality and the sciences of memory*. Princeton: Princeton University Press.
Halbwachs, M. (1941). *On collective memory* (Ed. & Trans. Lewis A Coser). Chicago, IL: University of Chicago Press.
Kirsch, D. (2009). Problem solving and situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 264–306). New York: Cambridge University Press.
Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. London: Penguin.
Levy, N. (2007a). Rethinking neuroethics in the light of the extended mind thesis. *The American Journal of Bioethics, 7*(9), 3–11.
Levy, N. (2007b). *Neuroethics: Philosophical challenges for the 21st century*. New York: Cambridge University Press.
Lindemann, H. (2010). Holding one another (well, wrongly, clumsily) in a time of dementia. In E. F. Kittay & L. Carlson (Eds.), *Cognitive disability and its challenge to moral philosophy* (pp. 161–168). Malden: Blackwell.
Locke, J. (1690). *An essay concerning human understanding*. New York: Oxford University Press.
MacIntyre, A. (1984). *After virtue: A study in moral theory*. Notre Dame: University of Notre Dame Press.
Marks, C. E. (1981). *Commissurotomy and consciousness and unity of mind*. Cambridge: MIT Press.
Neisser, U. (1967). *Cognitive psychology*. New York: Appleton.
Nozick, R. (1981). *Philosophical explanations*. Cambridge: The Belknap Press of Harvard University Press.
Olick, J. (2011). *The collective memory reader*. New York: Oxford University Press.
Olson, E. T. (2003). An Argument for Animalism. Author manuscript available at: http://eprints.whiterose.ac.uk/archive/00000734/ Accessed 2 September 2009. Also published in: Olson, E.T. (2003). An argument for animalism. In: Martin, R. and Barresi, J., (eds). *Personal identity*. Blackwell readings in philosophy (11). Oxford: Blackwell, pp. 318–334.
Olson, E. T. (1997). *The human animal: Personal identity without psychology*. New York: Oxford University Press.
Puccetti, R. (1981). The case for mental duality: Evidence from split-brain data and other considerations. *The Behavioral and Brain Sciences, 4*, 93–123.
Puccetti, R. (1973a). Brain bisection and personal identity. *The British Journal for the Philosophy of Science, 24*, 339–355.
Puccetti, R. (1973b). Multiple identity. *The Personalist, 54*, 203–215.
Rupert, R. (2009). *Cognitive systems and the extended mind*. New York: Oxford University Press.
Sawyer, R. K., & Greeno, J. G. (2009). Situativity and learning. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 347–367). New York: Cambridge University Press.
Schacter, D. L. (2010). Implicit vs. explicit memory. In R. A. Wilson & F. Keil (Eds.), *The MIT encyclopedia of cognitive sciences* (pp. 394–395). Cambridge: MIT Press.
Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994*. Cambridge, MA: MIT Press.

Schneider, S. (2009). Mindscan: Transcending and enhancing the human brain. In S. Schneider (Ed.), *Science fiction and philosophy: From time travel to superintelligence*. Oxford: Wiley-Blackwell.
Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20*(11), 11–21.
Segal, G. (1989). Seeing what is not there. *Philosophical Review, 98*, 189–214.
Shapiro, L. (1993). Content, kinds, and individualism in Marr's theory of vision. *Philosophical Review, 102*, 489–513.
Shoemaker, D. (2012). Personal identity and ethics. In Edward N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/identity-ethics/.
Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*, 1–29.
Sperry, R. (1966). Brain bisection and mechanisms of consciousness. In J. Eccles (Ed.), *Brain and conscious experience* (pp. 298–313). New York: Springer.
Sperry, R. (1968a). Mental unity following surgical disconnection of the cerebral hemispheres. *Harvey Lectures, 62*, 714–722.
Sperry, R. (1968b). Hemisphere deconnection and unity in conscious awareness. *American Psychologist, 23*, 723–733.
Theiner, G. (2008). *From extended minds to group minds: Rethinking the boundaries of the mental*. Ph.D. thesis, Indiana University.
Theiner, G. (2013). Onwards and upwards with the extended mind: From individual to collective epistemic action. In L. R. Caporael, J. Griesemer, & W.C. Wimsatt (Eds.), *Developing scaffolds in evolution, culture, and cognition*. Cambridge, MA: MIT Press.
Tulving, E. (2010). Episodic vs. semantic memory. In R. A. Wilson & F. Keil (Eds.), *The MIT encyclopedia of cognitive sciences* (pp. 278–280). Cambridge: MIT Press.
Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon.
Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic.
Wheeler, M. (2010). In defense of extended functionalism. In R. Menary (Ed.), *The extended mind*. Cambridge, MA: MIT Press.
Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.
Wilson, R.A. (in press). Ten questions concerning extended cognition, in a special issue of *Philosophical psychology*, edited by T. Sturm and A. Estany.
Wilson, R. A. (2010). Extended vision. In N. Gangopadhyay, M. Madary, & F. Spicer (Eds.), *Perception, action and consciousness* (pp. 277–290). New York: Oxford University Press.
Wilson, R. A. (2005). Collective memory, group minds, and the extended mind thesis. *Cognitive Processing, 6*, 227–236.
Wilson, R. A. (2004). *Boundaries of the mind: The individual in the fragile sciences*. New York: Cambridge University Press.
Wilson, R. A. (1995). *Cartesian psychology and physical minds: Individualism and the sciences of the mind*. New York: Cambridge University Press.
Wilson, R. A. (1994). Wide computationalism. *Mind, 103*, 351–372.
Wilson, R. A., & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 55–77). New York: Cambridge University Press.

Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System

27

Jennifer Chandler

Contents

Abstract

The objective of this chapter is to consider how emerging neuroscience might affect the way that the concept of personal identity is understood and used in the law. This topic is explored using two well-established theoretical approaches to the concept of personal identity. One approach considers the physical and/or psychological criteria for establishing the boundaries of one single personal identity at a given time (synchronic numerical personal identity) or the continuity of one personal identity over time (diachronic numerical personal identity). Another approach conceives of personal identity as "narrative identity" or the self-conception that a person creates from the sum of their experiences, values, and psychological attributes.

J. Chandler
Faculty of Law, Common Law, University of Ottawa, Ottawa, ON, Canada
e-mail: chandler@uottawa.ca; Jennifer.Chandler@uOttawa.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_115,

A concern with what makes two apparent beings the same person at one point in time (synchronic identity) brings into focus questions about how the law should respond to cases of accused persons with dissociative identity disorder. Neuroimaging and psychological research into dissociative identity disorder may one day alter the conceptualization of this disorder in ways that may affect the legal response to determining criminal responsibility in such cases. Meanwhile, a concern with changes in the "self" brings into focus a range of legal issues posed by emerging neurological interventions. The chapter offers three illustrative examples drawn from criminal and civil law: (1) What are the limits on legally coerced consent to "self"-changing rehabilitative brain interventions in the criminal context? (2) Should there be an expanded risk disclosure discussion during the informed consent process for medical treatment that may alter the "self"? (3) Who might be legally responsible for illegal behavior committed following "self"-changing brain interventions?

Introduction

The objective of this chapter is to consider how emerging neuroscience might affect the way that the concept of personal identity is understood and used in the law.[1] The law does not employ a single unified concept or theory of personal identity. Instead, personal identity and related concepts such as personhood are used in a variety of legal contexts and for a variety of purposes such as the recognition of legal rights and duties, the assignment of legal liability, and the creation of property rights (see, e.g., Radin 1982, discussing the personhood theory of property rights).

This chapter links the law to two well-established theoretical approaches to personal identity. First, the question of numerical personal identity, or the conditions that make one person the same person at a different time (diachronic identity) or that establish that two beings are the same person at one point in time (synchronic identity), is a familiar topic in philosophy (Olson 2003; DeGrazia 2005). Since only the person who committed an alleged crime may be held criminally responsible for it, the appropriate criterion for establishing that an accused is the person who committed an allegedly criminal act is important. In most cases, this is a practical problem flowing from the frailties of the identification evidence available, rather than a philosophical problem of numerical identity. However, as discussed below, the example of dissociative identity disorder has forced courts to grapple with the possibility that multiple legally distinct identities might exist within one body. Continuing neuroscientific and psychological research into this disorder may come to influence judicial reasoning in these cases.

[1]The legal principles and cases used in this chapter to illustrate the intersection of concepts of personal identity, neuroscientific knowledge, and the law are predominantly Canadian.Similar issues will confront other legal jurisdictions, where the reasoning and principles may be different.

The second theoretical approach reflected in this chapter is that of "narrative identity." This approach is concerned with a person's answer to the question "who am I?" This is a question that could produce many answers, and there are many elements that are conceivably relevant to the formation of one's self-conception (Merkel et al. 2007). Schermer (2011) describes narrative identity as a "person's self-conception, his biography, values, and roles as well as his psychological characteristics and style." Thus, one's behavior (both past actions and plans for the future), personality, values, and social roles and relationships are knitted together into an intelligible account of oneself over time that constitutes one's narrative identity. Theories of narrative identity generally recognize that the story of the self emerges within a social context such that "we are never more (and sometimes less) than the co-authors of our own narratives" (Schechtman 2011, citing MacIntyre 1984; Baylis 2011).

The breadth of inquiry permitted by narrative theories of personal identity is useful for this chapter as it allows for a focus on a wider range of legal preoccupations that might ultimately be affected by neuroscience. Furthermore, the neuroethics literature frequently raises issues related to the impact on self-conception of neuroscience, neuroessentialism, and emerging methods of manipulating brain function. This chapter thus links to that literature in identifying legal dimensions of the topics of interest to neuroethicists.

Accordingly, the second part of this chapter addresses three particular instances in which the law, neuroscience, and narrative identity come together. Each of these instances involves a change in mental attributes (cognitive functions, personality traits) that are a fundamental part of the answer to the question "who am I," and in each case, a legal question is raised. The first question is how we should distinguish between those brain interventions that a convicted criminal could be legally coerced into accepting in order to avoid imprisonment and those that could not legitimately be offered in this way. The second issue relates to the law of informed consent to medical treatment and asks whether the scope of risk disclosure should be broadened for brain interventions to include a range of social and legal consequences of treatment. The third issue relates to who might be responsible for harm caused by a patient who has undergone a brain intervention such as deep brain stimulation. Together, these three topics are meant to illustrate the ways in which neuroscience, particularly interventions that affect the "self," poses challenges for the law.

Personal Identity and Responsibility: "Who Did It?"

From the legal perspective, personal identity is of central importance in ascertaining who is to be held responsible for a crime. Put simply, the objective is to determine "who did it?" Instances of wrongful conviction – where the person convicted and punished is later found not to have been the person who committed the crime – are recognized as miscarriages of justice. The issue of numerical

identity – are "this and that…one thing, rather than two" (Olson 2003) – is a familiar topic in the philosophy of personal identity. One may ask this question as an issue of persistence over time, in which case the question is under what conditions can we say that a person at one time is the same as one that exists at another time? Theories of personal identity tend to emphasize the importance of some degree of physical continuity and/or psychological continuity as the criteria that establish personal identity over time (Olson 2003). One may also ask the question of synchronic numerical identity, as in the case of dissociative identity disorder, in which case the problem is to understand what establishes the boundaries of a distinct personal identity at one point in time.

In criminal cases, a perpetrator may be identified on the basis of physical attributes (e.g., through eyewitness reports, fingerprints, or DNA), as well as using evidence such as opportunity or motive. Since a satisfactory level of psychological and physical continuity over time is the norm for human beings, the law rarely has to take a position on which of the two – mental attributes or physical attributes – establishes that the accused is actually the person who committed the crime. The familiar philosophical thought experiment of tracking personal identity where the cerebrum is removed from one person and transplanted to another is not the type of complication that the law must address. Rather, concerns related to identity flow more from the practical problem of the quality of the evidence of physical identity, as wrongful convictions due to the frailty of eyewitness evidence demonstrate.

However, in some cases, the law is forced to grapple with uncertainties about whether the accused is the "same person" who committed the alleged crime, notwithstanding that personal identity seems to have been established on a physical basis. The issue of dissociative identity disorder has arisen in a range of cases where perplexed courts have had to decide whether to believe the claim of the "host" identity that an "alter" identity committed the allegedly criminal act. Where they do believe it, they must then decide whose mental state and capacity must be examined to determine the existence of mens rea, and how one can punish just one identity without simultaneously punishing "innocent" identities.

Dissociative identity disorder is not the only context in which dissociation must be addressed by the courts. The law also addresses criminal acts committed during a state of dissociation (sometimes called "automatism" in legal terminology). In these cases, the law typically continues to assume that the dissociated person is still the *same* person while committing the offence, and instead approaches the matter as a question of whether the requisite mental state for responsibility was present at the time. Isolated incidents where there is no underlying mental disorder that suggests a risk of recurrence may be treated as aberrations leading to an acquittal (e.g., *R. v. Bohak* 2004). Most often, however, courts will find the accused not criminally responsible due to mental disorder (formerly, the defense of insanity) (*R. v. Luedecke* 2008; *R. v. Stone* 1999).

The more extreme case of dissociative identity disorder, in which well-developed alternate personalities are said to inhabit one physical person, comes closer to obliging a court to grapple with the proper criterion for personal identity.

As one Canadian judge put it when sentencing a man who had pleaded guilty, but whose crime seemed to have been committed by an alternate personality,

> The problem is that the "person" before me is not the "person" who committed the crime nor is it the "person" who will serve the sentence. Physically, the appropriate individual is present in Court, but the guiding, controlling force which perpetrated the crimes is deep within the physical body of Stephen and waits there, unaffected by any conventional sentence I might impose. . .Though the current state of the law requires me to sentence the physical being who committed the offenses, I am permitted to take into account as I craft my sentence that, in essence, another being orchestrated the events. (*R. v. O'Flaherty* [1992] A.J. No. 96)

Courts have responded to claims of dissociative identity disorder in various ways, sometimes refusing to believe the claims. Where the claims are believed, some courts focus on the state of mind of the identity in charge at the time of the allegedly criminal act and hold the accused guilty if that identity had the requisite mens rea. Others focus on the mental state of the "host" or dominant identity, with attention to whether the "host" can be said to have been incapable due to mental disorder (Crego 2000). Although fairly rare, cases in which dissociative identity disorder are raised continue to surface (e.g., *State v. Dumas* 2012).

Dissociative identity disorder remains a contested and poorly understood condition (Boysen and Van Bergen 2013; Boysen 2011). Psychological and neuroscientific research is proceeding and may eventually affect legal responses to the disorder. For example, Reinders (2008) speculates that structural and functional neuroimaging may one day help in distinguishing between real and feigned cases of dissociative identity disorder, although she cautions that existing research is inadequate for forensic use. However, even if such evidence confirms that a case of dissociative identity disorder is genuine rather than feigned, the law must still address the problem of whether the accused person is to be considered legally responsible. The legal threshold for responsibility looks not at medical diagnosis but at the capacity to understand the nature and quality of the acts in question and to know that they are wrong. As a result, the problem of determining actual capacity at the time of the crime will remain even if neuroimaging one day is able to support a diagnosis of dissociative identity disorder.

Another possible effect of ongoing research is suggested by studies of the extent to which alternate identities are truly separate. If it is shown that there is actually overlap in mental functions between identities, the argument for identity based on distinct psychological attributes (which is what remains to establish separate identities given that they share the same body) may be weakened. In other words, the situation begins to look more like one of a single disordered identity than of multiple distinct identities within one body, each of which can be assessed independently for mental competence and criminal responsibility.

One line of research focuses on the extent to which one identity remembers words presented to another despite claimed inter-identity amnesia. Huntjens et al. (2007) conclude that there is evidence of the transfer of information between identities in dissociative identity disorder and that "[t]hese findings strikingly contrast with the patient's subjective reports of amnesia for the task performed

and material learned by the learning identity." They further suggest that studies of this type may be important for the conceptualization of dissociative identity disorder in the future, with the disorder being a disturbance in "metamemory functioning" or knowledge, beliefs, and feelings about memory, rather than an actual fragmentation of memory. Evidence of this type might encourage the law to make a global assessment of an accused person's mental state, rather than, for example, focusing on the competence of the identity "in control" at the time of the alleged crime.

Brain Interventions that Affect Narrative Identity: "Who Am I?"

Discussions of personal identity often arise in neuroethics in relation to interventions in the brain that may affect mental attributes important to the self, including cognitive functions and personality attributes. Two key themes are the evaluation of the change in the self (i.e., on what basis might we judge a change to be good or bad) and autonomy in relation to a change in the self (i.e., what are the conditions for an autonomous decision to change the self, and when is it acceptable to coerce a change in the self). A third issue, which arises less frequently in neuroethics, is the question of responsibility for behavior clearly associated with a major change in personality following a brain intervention.

This section selects three legal questions relating to brain interventions that may alter the mental functions and characteristics important to personal identity. The first issue emerges from criminal and human rights law and has to do with the limits on legally coerced consent to rehabilitative treatment in the criminal context. The second has to do with the civil legal question of the scope of risk disclosure required for informed consent to medical treatment in the context of emerging brain interventions where a range of nonmedical (i.e., social or legal) consequences may flow from treatment. Finally, the third question is who might be responsible for harm caused by a patient who has undergone a brain intervention such as deep brain stimulation. Together, these three topics are meant to illustrate the ways in which neuroscience, particularly interventions that affect the "self," poses challenges for the law.

What are the Limits on Legally Coerced Consent to Identity-Changing Brain Interventions in Criminal Offenders?

The law pressures convicted criminal offenders to accept treatment at several stages of an offender's transit through the criminal justice system. For example, acceptance of treatment allows some offenders to avoid incarceration through diversion to specialized drug treatment courts (e.g., the Toronto Drug Treatment Court). Cooperation with treatment is also considered at sentencing and in parole decision-making. Sentencing judges tend to view acceptance of treatment favorably because it is thought to demonstrate better prospects for rehabilitation. Judges also sometimes

view offenders who accept treatment as showing remorse and taking responsibility for their crimes, both of which can be mitigating circumstances at sentencing. Acceptance of treatment is particularly important for serious repeat offenders who are at risk of indefinite preventive detention as "dangerous offenders" under the Canadian Criminal Code. Those whose riskiness is deemed to be controllable in the community (e.g., through medical treatment) may be released under long-term supervision orders following completion of their prison terms. A range of biological interventions that arguably change important aspects of personality and behavior are available, such as the widely used antiandrogen drug therapy ("chemical castration") for sex offenders (Grubin and Beech 2010; *Deacon* 2006).

As we accumulate more knowledge about neurobiological predispositions to criminal behavior and we develop a broader array of potential techniques to intervene in the brain, the issue of what types of treatments an offender can be legally coerced to accept rises in importance. The temporary resurgence of psychosurgery for drug addiction in China and Russia (Lu et al. 2009; Hall 2006; Stelten et al. 2008), as well as the possibilities of deep brain stimulation for addiction (Luigjes et al. 2012) and psychiatric indications, demonstrates the importance of this question.

The answer to the question of whether an offender may legitimately be asked to consent to a particular brain intervention in exchange for a legal advantage depends upon how one categorizes the offered alternative from the perspective of the philosophical justifications of punishment. From the retributivist perspective, punishment is the infliction of deserved harm on an offender (see, e.g., Honderich 2006, for a discussion of this proposition). The extent and type of harm is subject to the limits posed by the human right not to be subjected to "cruel and unusual punishment" (e.g., UN General Assembly, 1948, article 5; Canadian Charter of Rights and Freedoms 1982, Sect. 12). Various consequentialist justifications of punishment also exist, one of which is that punishment should incapacitate offenders in order to stop them from offending again. From this perspective, a biological intervention might be understood not as retributive harm but as a way to protect the public by moving the prison walls into the body – a kind of "somatic regulation" to replace supervision or incarceration (Vrecko 2010). This vision of criminal justice as "risk management" is concerned primarily with the effective protection of the public rather than with the well-being of the offender. Finally, a biological intervention may also be viewed as rehabilitative, the chief focus of which is to change the offender for his or her presumed benefit as well as the general social benefit. From a practical perspective, rehabilitation is likely the appropriate theoretical frame in which to consider a biological intervention since the participation of physicians is likely to be needed, and their ethical participation requires that the treatment benefit the offender. Although physicians practicing in the correctional context face challenging problems of dual loyalty and conflicting obligations to patients, other inmates, and society (Konrad and Völlm 2010, they are bound by ethical obligations of beneficence and non-maleficence to their patients (UN General Assembly 1982, Principle 3; Elger 2008). From this perspective, the limits on acceptable legally coerced biological interventions are set

by whether the offered intervention would produce a net benefit or harm to the offender – a calculation that is far from easy to make.

Whether a biological intervention actually inflicts net benefit or harm on the offender depends upon whether the personality and behavioral changes produced by the intervention (e.g., reducing the risk of future criminality) are more valuable to the offender than what has been lost as a result of the treatment. Whether this is the case is a complex question involving the effects of the intervention on physical and mental health and functioning, the effects on personal identity and autonomy, and the issue of whether there is inherent value in living in a way that conforms to social norms.

The safety and effects of biological interventions on physical health and mental functioning are evidently key components of assessing whether they are harmful or beneficial. This consideration is important given that there is a risk that societies may be tempted to use unproven treatments prematurely on offenders, who are an unpopular social group (Greely 2012).

The impact on personal identity of biological interventions designed to modify personality and behavior has also been raised as a key concern (Levy 2007; Shaw 2012; Vincent 2012). As narrative personal identity is always changing over a lifetime, it is implausible to suggest that change is in itself harmful. Instead, the evaluation must focus on a particular change in the context of a particular life. Relevant features of the change in personal identity that might affect the evaluation include its nature, magnitude, abruptness, permanence, and the degree to which the changed person is aware of and endorses the change as reflecting his or her authentic self. One concern with direct biological treatments such as interventions in the brain is that, unlike psychological "talk therapies" that rely at least in part on appeals to the rationality of the offender, biological interventions bypass the rational mind, working changes that the offender does not have the opportunity to resist or accept (Levy 2007; Shaw 2012; Bublitz and Merkel 2009; Vincent 2012; Bomann-Larsen 2011; Greely 2012). At the same time, some biological interventions may restore a capacity for self-awareness and critical reflection (Shaw 2012) or may create a subjective sense of feeling "more like oneself" (Bolt and Schermer 2009). In these cases, biological interventions are arguably beneficial rather than harmful to personal identity.

Autonomy is another value that may be affected by biological interventions that affect personality and behavior. Legal coercion to accept any type of treatment is an affront to autonomy. Medical ethics will need to determine whether the benefits of treatment outweigh the harm to autonomy of legally coerced consent. Thus, we leave behind the fact that any coerced treatment harms the autonomy of a competent person. Instead, we focus on the effect of biological interventions on autonomy in order to determine whether the interventions produce a net benefit or harm to the offender. Put another way, the primary question here is whether the offender's capacity for autonomous decision-making is increased or decreased as a result of the intervention. Frankfurt's (1971) distinction between higher- and lower-order desires is helpful here, as is the story of Odysseus and the Sirens. Sometimes an external means to control desires that are rejected by the self is a means to allow the

"real" self to flourish more fully. For example, some sex offenders report benefit from "no longer being preoccupied by sexual thoughts or dominated by sexual drive" so that they are able to participate in psychological treatment without distraction (Grubin and Beech 2010; BBC News 2009). When someone rejects an internal compulsion as a limit on their autonomy, a treatment that removes this impulse or enhances self-control may thus support their autonomy. Things are more complex when a person does not reject the lower-order desire at the outset but comes to reject it during treatment, so that the future person might be said to have increased autonomy while the autonomy of the present person is eroded.

As for the issue of whether there is inherent value in living in a way that conforms to social norms, there is undoubted value in avoiding social condemnation and punishment. We should not underestimate the value of avoiding "liberty-depriving 'total institutions' [such as prisons]. . .that sever ties to outside families, friends, jobs and communities" (Vrecko 2010). However, it is dangerous to assume that the ways of living promoted by social norms are inherently valuable or that criminalized ways of living are without value. The history of legally coerced "curative treatment" of homosexuality illustrates this point (King et al. 2004; Joyce 2009). Rehabilitation – imposed paternalistically for an offender's own good – has also been criticized as highly dangerous because it is sheltered by an ideology of humanity and beneficence that allows for just about anything to be done as long as experts regard it as beneficial to the offender (Lewis 1970). Furthermore, it is not always clear that the traits that biological interventions might target are always maladaptive (Hörstkotter et al. 2012).

In addition to these potential benefits and harms to the offender of legally coerced biological interventions, there are possible consequences for the broader society. Some have argued that rehabilitation that changes behavior and attitudes removes the very individuals best suited to argue for and represent dissident values, impeding possibly salutary evolution of morality and law (Durkheim 1982; Shaw 2012). Another possible harm to society is that we defuse one of the pressures for society to confront other criminogenic factors (e.g., social inequality) that may be challenging but important to address. Finally, legally coerced treatment creates a practical problem of equitable resource distribution. In the context of specialized courts such as drug treatment or mental health courts, it has been observed that the result of court-ordered treatment is that offenders jump the treatment queue ahead of non-offenders (Seddon 2007).

In summary, as knowledge accumulates in behavioral neuroscience and we develop new techniques of intervention in the brain to change behavior, we will need to confront the question of the proper boundaries on the coerced rehabilitative treatment of criminal offenders. Given the necessary participation of medical practitioners bound by medical ethical obligations, only treatments that are of net benefit to offenders can be coercively offered to them. Human rights guarantees, particularly the right to be free from cruel and unusual punishment, may also buttress this position. However, it is very difficult to determine benefit and harm in identity-altering treatments. One's identity changes over time, frequently in ways and as a result of experiences we have not chosen. Nonetheless, the idea of abrupt

treatment-driven changes that bypass the rational mind are disturbing to narrative identity – understood as a self-created story of one's life and an answer to the question "who am I?"

Should the Scope of Risk Disclosure Required for Informed Consent to Identity-Changing Brain Interventions be Broadened to Encompass Potential Legal Consequences?

Memory is one of the mental functions cited as fundamental to personal identity, and alteration in memories may thus be perceived as affecting identity. In recent years, an experimental treatment for posttraumatic stress disorder (PTSD) that has come to be known as "memory dampening" has attracted ethical attention. Among the concerns raised have been questions of its effects on personal identity and authenticity (see the summary in Chandler et al. submitted, 2013).

This novel pharmacological intervention is currently being explored in clinical trials and uses the beta-blocker, propranolol, to treat PTSD. Much of the ethical discussion of this treatment was based on the assumption that propranolol must be administered at or soon after the traumatic experience. This raised a concern, among other issues, about the unnecessary treatment of people who might not go on to develop PTSD which might expose them to side effects, waste health-care resources, and possibly deny them the opportunity for the positive psychological experience of "posttraumatic growth" (Warnick 2007; Calhoun and Tedeschi 2006). However, it now appears that propranolol may be administered after PTSD develops to disrupt the reconsolidation of traumatic memories (Brunet et al 2011, 2008; Menzies 2012, 2009). The evidence is mixed on whether this is an effective PTSD treatment, and also on the effects of the treatment on a patient's recall of factual details and emotional response to the traumatic memory. However, Menzies (2009) reports that his patients experienced "'fragmentation' of the memory and difficulty accessing it, minimal or absent distress when thinking about it and a feeling of emotional detachment, as if it were a normal non-traumatic memory or had happened to someone else." In a later larger study, Menzies (2012) reported that the treatment "diminished the integrity of these traumatic memories, resulting in a degree of amnesia for the traumatic event." Additional studies are required to understand the effects of propranolol on traumatic memories. However, if it is true that this treatment does produce fragmented memories, this might undermine a witness's ability to testify persuasively in court. Similarly, an oddly detached style of recounting a horrible experience might also undermine a witness's credibility. Since many traumatic experiences might give rise to criminal or civil legal proceedings, this treatment might thus have consequences for the patient's ability to testify effectively. Should this possible consequence of treatment form part of the risk disclosure required for a patient to give informed consent (Chandler et al. submitted, 2013)?

With narrow exceptions, medical law and ethics require that informed consent be given for medical treatment, either directly by the patient or by the patient's

substitute decision-maker if the patient is incompetent. Canadian law indicates that physicians must disclose the "material risks" of a proposed treatment, of alternative treatments, and of nontreatment. There is some vagueness in what counts as a material risk, although severity and likelihood are key criteria in determining whether a potential adverse outcome must be disclosed. Even if the ordinary person would not regard a given risk as material, it may need to be disclosed if a patient asks questions that put a physician on notice that a patient has particular sensitivities rendering the risk material to that patient (Picard and Robertson 2007). Sometimes a patient may describe a specific social factor (e.g., a concern with scarring flowing from the desire to conceal her contraceptive sterilization from her family, *Videto v. Kennedy* 1981) that makes her unusually sensitive to particular medical risks. Note, however, that the focus is still on identifying which medical risks must be disclosed, rather than on identifying and disclosing the range of indirect social consequences that must also be identified as risks of treatment.

Where standard medical treatments are at issue, the law of informed consent typically focuses on the obligation to disclose information about the direct medical effects of treatment (rather than their indirect social, economic, or legal significance) and, to some extent, also focuses on the obligation to disclose information about the physician (Picard and Robertson 2007; Borden Ladner Gervais 2000; Berg et al. 2001). In the context of this discussion, however, the question is whether the broader social consequences of the medical effects of the proposed therapy should be discussed with patients. One assumes that the direct effects of propranolol treatment on a patient's memories for a traumatic event would be discussed with the patient as this is the therapeutic objective. The remaining question is whether the indirect legal consequences of those direct therapeutic effects should also be discussed. There are reasons to think that such indirect consequences should be part of the risk disclosure process in the context of emerging treatments. For example, ethicists and lawyers have called for an obligation to disclose to patients the broader social, legal, and economic risks of learning genetic information about themselves in the context of genetic screening (Offitt and Thom 2007). Similar calls have been made to inform prospective research participants that one of the risks of participation is that "incidental findings" may be discovered in the process of genetic or imaging research and that this information may affect insurability or may disrupt family relationships (e.g., discovery of misattributed paternity) (Wolf et al. 2008; Smith Apold and Downie 2011). This obligation is arguably inherent in the major Canadian research ethics code, which requires researchers to have a plan approved by the ethics review board for the disclosure to participants of incidental findings (findings with "significant welfare implications. . .whether health-related, psychological or social") (TCPS 2 2010). Disclosure to prospective participants that incidental findings with significant welfare implications are anticipated ought to fit within the obligation to discuss the risks of participation. Some research ethics codes already address the issue of the disclosure to participants of the social, legal, and economic risks related to the genetic information that may be revealed during research (e.g., TCPS 2 2010; NHGRI 2012). While research is not the same as

therapeutic treatment, and so the scope of risk disclosure differs, perhaps novel treatments ought also to entail a broader scope of risk disclosure (i.e., that extends to social, economic, or legal consequences of the anticipated medical effects) than appears to be the case for well-established medical treatments.

The justification for expanded risk disclosure for emerging and experimental treatments is that these are the situations in which patients may be least capable of anticipating the indirect social, legal, or economic consequences of proposed treatments. Should the scope of risk disclosure be broadened to include such consequences where an emerging treatment may affect the fundamental building blocks of the self (i.e., cognitive functions and personality traits)? Schmitz-Luhn et al. (2012) raise the possibility that the legal requirements for risk disclosure and informed consent may be broader than normal in the context of deep brain stimulation because of the possibility of social consequences and personality change. They explore whether it is necessary to go beyond mentioning the possibility of these effects and to provide counseling to help the patient understand the potential psychosocial consequences in his or her particular case. In the case of memory dampening, should the risks of adverse legal consequences of treatment be included as part of the risk disclosure by physicians? It is true that physicians are not legally trained and may not be well placed to anticipate and advise on these matters. On the other hand, since many PTSD patients have suffered traumas that may foreseeably result in lawsuits or prosecutions, perhaps it is not unreasonable to expect physicians to be aware of and mention the issue, particularly since the timing of the propranolol treatment could be varied for someone who will likely serve as a witness.

Novel medical treatments often pose problems of this type, as the broader ramifications of a new treatment may be unclear to patients and physicians alike. As a novel treatment and its broader consequences become more familiar, it may be reasonable to expect patients to identify and weigh for themselves the significance of these broader consequences. One possible solution in the meantime is to make more use of practice guidelines for emerging treatments, some of which may direct physicians to raise broader social and legal ramifications with their patients. This is not unprecedented. For example, some clinical practice guidelines for assisted reproductive medicine include counseling about the psychosocial and legal aspects of using donor gametes (NHMRC 2007).

Who is Legally Responsible for Harmful Behavior that Flows from Brain Interventions such as Deep Brain Stimulation?

Brain interventions such as deep brain stimulation (DBS) occasionally produce dramatic unintended changes in mood, cognition, and behavior and may thus deeply influence a patient's personal identity. Reports from some DBS patients suggest that these changes can destabilize the subjective sense of self. Some patients are reported to have said "I don't feel like myself anymore," or "I feel like a robot" (Schüpbach et al. 2006). In a minority of patients, more significant psychiatric and behavioral problems may ensue, including severe depression,

hypersexuality, aggression, hypomania, and mania (Clausen 2010). In some cases, the person is so different that "out of character" criminality results. In one case, a hypomanic patient faced prosecution after he broke into a car in a crowded street (Mandat et al. 2006). Several interesting questions related to legal responsibility arise in this context. Are patients, and possibly their physicians, legally liable for harmful acts that a patient commits while under the influence of the stimulation?

A DBS patient who breaks the criminal law while under the influence of brain stimulation is likely to be held criminally responsible unless the impact of the stimulation is so severe that the patient meets the requirements for the defense of insanity (in Canada, "not criminally responsible by reason of mental disorder"). As for civil liability for actions while under the influence of DBS, legal systems vary in whether mental capacity is a requirement or not for liability in tort. In the United States, for example, people with mental illness or disability are usually responsible for both intentional torts and negligence (Dobbs 2000). In Canada, there is some uncertainty in the law, but some courts have excused mentally ill persons from liability for intentional torts and negligence where they are incapable of appreciating the nature and quality of their acts or of appreciating and discharging the duty of care they owed (Linden and Feldthusen 2006). However, in Canada, it is clear that a person who ought reasonably to foresee their own mental incapacity and who does nothing to avoid causing harm while in that incapable state will be held liable in negligence for harms caused while incapable. As a result, a DBS patient who is warned of the possibility of hypomania and fails to take steps to avoid causing harm when the stimulator is turned on may be held liable in negligence for damages caused (e.g., for dangerous driving).

As for the tort liability of physicians for harms caused by DBS patients as a result of the stimulation, it is useful to distinguish potential liability to the patient where the patient harms himself or herself and to third parties harmed by the patient. Starting with the patient who has harmed himself or herself, a physician could be liable to the DBS patient in two ways: First, the administration of DBS might fall below the requisite standard of care. This could occur, for example, if the particular patient was not suitable for DBS in the first place (e.g., the risks, including behavioral risks, outweighed the benefits for the patient), if the surgery was performed in a way that negligently enhanced the risk of harmful behavioral side effects, or if the follow-up monitoring and care were performed negligently in a way that failed to deal with harmful behavioral side effects. Second, a physician may be liable in negligence for failing to disclose material risks such as harmful behavioral consequences. Liability for inadequate risk disclosure would depend upon a showing that the lack of disclosure affected the decision. In Canada, this requires the patient to show that a reasonable person in the patient's position would not have gone ahead with the treatment if proper disclosure of the risks had been made (Picard and Robertson 2007).

The analysis is more challenging in the context of a physician's possible responsibilities to third parties. Where a defendant has some hand in "creating" a risky individual (e.g., by serving alcoholic drinks to the point of intoxication in

a bar, *Stewart v. Pettie* 1995; or by allowing a patient to drive home after an emotionally upsetting outpatient procedure, *MacPhail v. Desrosiers* 1998) or has a duty of supervision and control over a risky person (e.g., responsibility for an institutionalized patient with a mental illness, Picard and Robertson 2007), negligence law may impose liability on the defendant for harms caused by the risky individual to third parties. However, liability will follow only if the defendant fails to take reasonable steps to prevent the risky individual from causing harm.

In addition, physicians are sometimes held responsible if they fail to warn others (e.g., the police or third parties at risk) where a patient poses a reasonably foreseeable serious risk to third parties (e.g., *Tarasoff v. Regents of the University of California* 1976). The duty to warn collides with the duty of confidentiality to the patient, and so usually arises only where the danger is serious and imminent. Separate statutory notification requirements may also exist, as with the obligation in Ontario to notify the Ministry of Transportation of patients who are unfit to drive. The failure to comply with this statutory notification requirement may be used in a subsequent lawsuit against physicians by those injured by the patient (Picard and Robertson 2007).

As a result, it is conceivable that physicians may have some responsibility toward third parties to take reasonable steps to address serious behavioral side effects in their DBS patients. The outcome of such cases will depend very much on the specific facts of the case, including the seriousness of the behavioral effects in the particular patient, the foreseeability of harm caused by the patient, and the nature of the attempts to reduce the risk.

Conclusion

The law, neuroscience, and the concept of personal identity intersect in a multitude of interesting ways. This chapter has sought to organize this "area of intersection" using two well-established theoretical approaches to the concept of personal identity.

One of the traditional philosophical questions relating to personal identity is that of numerical personal identity, or the task of understanding what makes a person at one time the same person at another time and also what establishes the boundaries of one single person at one particular time. This latter concern, of synchronic numerical personal identity, is a central issue for the law in addressing dissociative identity disorder. In these cases, the law must decide what to do where an accused person claims that a "host" identity lacked knowledge or control over the actions of an "alter" identity. Neuroimaging research into dissociative identity disorder may one day supply additional diagnostic evidence and so may help to distinguish between genuine and feigned cases, but this alone will not resolve the difficult question of establishing legal responsibility in non-feigned cases of dissociative identity disorder. Another line of research into the overlap in mental functions of the various identities might have an effect on the law by undermining the idea that the identities are truly distinct and can be individually assessed for capacity and criminal

responsibility. This would tend to support a legal approach that looks at the accused person as having one disordered identity and that therefore applies the standard legal test for mental capacity to that single person, taking into consideration the extent and severity of the dissociative symptoms in assessing the accused's capacity at the time of the offence.

The chapter also adopts the broader vision of personal identity reflected in the neuroethics literature's concerns with the consequences for the "self" of various brain interventions. These concerns flow from a conceptualization of personal identity in the form of narrative identity or the self-conception that a person creates from the sum total of their experiences, values, and psychological attributes. A focus on changes in psychological attributes, including memories and personality traits, leads us to consider another set of legal issues posed by emerging neurological interventions. The chapter offers three questions of this type. First, we must consider the limits on the extent to which the law may coerce a person to accept "self"-changing rehabilitative brain interventions in the criminal context. This is a complex question that relies in part on understanding when and why changes to personal identity are beneficial or harmful. Second, should there be an expanded risk disclosure discussion during the informed consent process for medical treatment that may alter the "self" (e.g., self-constituting attributes such as memories)? It seems reasonable that as we embark on novel interventions of this type, an expanded discussion of the broader implications of those alterations is in order, and there are ways that this might be done without unfairly burdening physicians with unrealistic expectations that they anticipate the broad and sometimes subtle ramifications of such treatments. Third, brain interventions that cause changes in personality and behavior raise the challenging questions of responsibility where that behavior is potentially criminal.

Acknowledgments Many thanks to Jocelyn Downie, Michael Hadskis, and Francoise Baylis for their most helpful comments on a draft of this chapter. All weaknesses or errors remain the sole responsibility of the author.

Cross-References

- ► Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
- ► Dissociative Identity Disorder and Narrative
- ► Drug Addiction and Criminal Responsibility
- ► Ethical Implications of Brain Stimulation
- ► Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
- ► Ethics in Psychiatry
- ► Ethical Issues in the Treatment of Addiction
- ► Impact of Brain Interventions on Personal Identity

▶ Neuroethics and Identity
▶ Neurolaw: Introduction
▶ Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry
▶ Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia
▶ What Is Addiction Neuroethics?

References

Baylis, F. (2011). The self in situ: A relational account of personal identity. In J. Downie & J. Llewellyn (Eds.), *Being relational* (pp. 109–131). Vancouver: UBC Press.

BBC News. 17 November 2009. *I was chemically castrated*. http://news.bbc.co.uk/2/hi/europe/8362605.stm.

Berg, J. W., Applebaum, S., Lidz, C. W., & Parker, L. S. (2001). *Informed consent: Legal theory and clinical practice*. New York: Oxford University Press.

Bolt, I., & Schermer, M. (2009). Psychopharmaceutical enhancers: Enhancing identity? *Neuroethics, 2*, 103–111.

Bomann-Larsen, L. (2011). Voluntary rehabilitation? On neurotechnological behavioral treatment, valid consent and (in)appropriate offers. *Neuroethics*. doi:10.1007/s12152-011-9105-9.

Borden Ladner Gervais LLP (2000). (looseleaf update 2011). *Canadian health law practice manual*. Markham: LexisNexis Canada.

Boysen, G. A. (2011). The scientific status of childhood dissociative identity disorder: A review of published research. *Psychotherapy and Psychosomatics, 80*, 329–334.

Boysen, G. A., & VanBergen, A. (2013). A review of published research on adult dissociative identity disorder 2000–2010. *Journal of Nervous and Mental Disease 201*(1), 5–11.

Brunet, A., et al. (2011). Trauma reactivation under the influence of propranolol decreases posttraumatic stress symptoms and disorder – 3 open-label trials. *Journal of Clinical Psychopharmacology, 31*(4), 547–550.

Brunet, A., et al. (2008). Effect of post-retrieval propranolol on psychophysiologic responding during subsequent script-driven traumatic imagery in post-traumatic stress disorder. *Journal of Psychiatric Research, 42*, 503–506.

Bublitz, J. C., & Merkel, R. (2009). Autonomy and authenticity of enhanced personality traits. *Bioethics, 23*(6), 360–374.

Calhoun, L. G., & Tedeschi, R. G. (Eds.). (2006). *Handbook of posttraumatic growth: Research and practice*. Mahwah: Lawrence Erlbaum.

Canadian Charter of Rights and Freedoms, Constitution Act, 1982, Schedule B to the Canada Act 1982 (UK), 1982, c. 11, available at http://www.canlii.org/en/ca/const/const1982.html.

Chandler, J.A., Mogyoros, A., Martin Rubio, T. and Racine, E. (submitted, 2013). Another look at the legal and ethical consequences of pharmacological memory dampening: The case of sexual assault.

Clausen, J. (2010). Ethical brain stimulation – neuroethics of deep brain stimulation in research and clinical practice. *European Journal of Neuroscience, 32*, 1152–1162.

Crego, M. E. (2000). One crime, many convicted: Dissociative identity disorder and the exclusion of expert testimony in State v. Greene. *Washington Law Review* 75, 911–939.

Deacon v. Canada (Attorney General). (2006) FCA 265.

DeGrazia, D. (2005). *Human identity and bioethics*. New York: Cambridge University Press.

Dobbs, D. B. (2000). *The Law of Torts. St*. Paul: West Group.

Durkheim, E. (1982). Rules for the distinction of the normal from the pathological. Trans. W.D. Halls. In Durkheim, E. (1982).*The rules of sociological method*. MacMillan Press Ltd. Reprinted with permission In M. Tonry (Ed.) (2011). *Why punish? How much?* (pp. 415–420) New York: Oxford University Press.

Elger, B. S. (2008). Medical ethics in correctional healthcare: An international comparison of guidelines. *The Journal of Clinical Ethics, 19*(3), 234–248.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.

Greely, H. (2012). Direct brain interventions to treat disfavored human behaviors: Ethical and social issues. *Clinical Pharmacology and Therapeutics, 91*(2), 163–165.

Grubin, D., & Beech, A. (2010). Chemical castration for sex offenders. *BMJ, 340*, c74.

Hall, W. (2006). Stereotactic neurosurgical treatment of addiction: Minimizing the chances of another 'great and desperate cure., *Addiction, 101*(1), 1–3.

Honderich, T. (2006). *Punishment: The supposed justifications revisited*. London: Pluto Press.

Hörstkotter, D., et al. (2012). "We are also normal humans, you know?" Views and attitudes of juvenile delinquents on antisocial behavior, neurobiology and prevention. *International Journal of Law and Psychiatry, 35*, 289–297.

Huntjens, R. J. C., et al. (2007). Memory transfer for emotionally valence words between identities in dissociative identity disorder. *Behaviour Research and Therapy 45*, 775–789.

Joyce, J. (2009). *Gay injustice "was widespread."* BBC News. 12 September 2009. http://news.bbc.co.uk/2/hi/uk_news/8251033.stm.

King, M., Smith, G., & Bartlett, A. (2004). Treatments of homosexuality in Britain since the 1950s – an oral history: The experience of professionals. *BMJ, 328*(7437), 427.

Konrad, N., & Völlm, B. (2010). Ethical issues in forensic and prison psychiatry. In H. Helmchen & N. Sartorius (Eds.), *Ethics in psychiatry, International Library of Ethics, Law, and the New Medicine 45*, (p. 363–380). Springer.

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Lewis, C. S. (1970). The humanitarian theory of punishment. Excerpt from Lewis, C.S. *God in the Dock*. C.S. Lewis Pte. Ltd. Reprinted in Tonry, M. ed. (2011). *Why punish? How much?* (pp. 91–96). New York: Oxford University Press.

Linden, A. M., & Feldthusen, B. (2006). *Canadian Tort law* (8th ed.). Markham: LexisNexis Canada.

Lu, L., Wang, X., & Kosten, T. R. (2009). Stereotactic neurosurgical treatment of drug addiction. *The American Journal of Drug and Alcohol Abuse, 35*(5), 391–393.

Luigjes, J., et al. (2012). Deep brain stimulation in addiction: A review of potential brain targets. *Molecular Psychiatry, 17*(6), 572–583.

MacIntyre, A. (1984). *After virtue* (2nd ed.). Notre Dame: Notre Dame University Press.

MacPhail v. Desrosiers, (1998). N.S.J. No. 353 (N.S.C.A.), varying MacPhail v. Desrosiers (1997). N.S.J. No. 562 (N.S.S.C.).

Mandat, T. S., Hurwitz, T., & Honey, C. R. (2006). Hypomania as an adverse effect of subthalamic nucleus stimulation: A report of two cases. *Acta Neurochirurgica (Wien), 148*(8), 895–898.

Menzies, R. P. D. (2009). Propranolol treatment of traumatic memories. *Advances in Psychiatric Treatment, 15*, 159–160.

Menzies, R. P. D. (2012). Propranolol, traumatic memories, and amnesia: A study of 36 cases. *The Journal of Clinical Psychiatry, 373*(1), 129–130.

Merkel, R., et al. (2007). *Intervening in the brain: Changing psyche and society* (Ethics of science and technology assessment, Vol. 29). Berlin: Springer.

NHGRI (National Human Genome Research Institute, National Institutes of Health, United States). (2012). *Informed consent elements tailored to genomics*. http://www.genome.gov/27026589.

NHMRC (National Health and Medical Research Council, Australia). (2007). *Ethical guidelines on the use of assisted reproductive technology in clinical practice and research*. http://www.nhmrc.gov.au/guidelines/publications/e78

Offitt, K., & Thom, P. (2007). Ethical and legal aspects of cancer genetic testing. *Seminars in Oncology, 34*(5), 435–443.

Olson, E. T. (2003). Personal identity. In S. P. Stich & T. A. Warfield (Eds.), *Blackwell guide to philosophy of mind* (pp. 352–368). Malden: Blackwell.

Picard, E. I., & Robertson, G. B. (2007). *Legal liability of doctors and hospitals in Canada* (4th ed.). Toronto: Thomson Carswell.

Radin, M. J. (1982). Property and personhood. *Stanford Law Review, 34*(5), 957–1015.

Reinders, A. A. T. S. (2008). Cross-examining dissociative identity disorder: Neuroimaging and etiology on trial. *Neurocase, 14*(1), 44–53.

R. v. Bohak, (2004). M.J. No. 172 (Manitoba Provincial Court).

R. v. Luedecke, (2008). ONCA 716.

R. v. O'Flaherty, (1992). A.J. No. 96 (Alta. Prov. Ct.).

R. v. Stone, (1999). 2 S.C.R. 290 (Supreme Court of Canada).

Schechtman, M. (2011). The narrative self. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 394–416). Oxford: Oxford University Press.

Schermer, M. (2011). Ethical issues in deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*(17), 1–5.

Schmitz-Luhn, B., Katenmeier, C., & Woopen, C. (2012). Law and ethics of deep brain stimulation. *International Journal of Law and Psychiatry, 35*, 130–136.

Schüpbach, M., et al. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*, 1811–1816.

Seddon, T. (2007). Coerced drug treatment in the criminal justice system: Conceptual, ethical and criminological issues. *Criminol and Crimal Justice, 7*, 269–285.

Shaw, E. (2012). Direct brain interventions and responsibility enhancement. *Criminal Law and Philosophy*. doi:10.1007/s11572-012-9152-2.

Smith Apold, V., & Downie, J. (2011). Bad news about bad news: The disclosure of risks to insurability in research consent processes. *Accountability in Research, 18*(1), 31–44.

State v. Dumas, (2012). Ohio App. LEXIS 56 (Court of Appeals of Ohio, 8th Appellate District).

Stelten, B.M.L. et al. (2008). The neurosurgical treatment of addiction. *Neurosurg. Focus* 25(1):E5.

Stewart v. Pettie, (1995). 1 S.C.R. 131.

Tarasoff v. Regents of the University of California, 551 P.2d 334 (Cal. 1976).

TCPS 2 (Tri-Council Policy Statement 2). (2010). Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada). Tri-Council Policy Statement. Ethical Conduct for Research Involving Humans. http://www.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf.

Toronto Drug Treatment Court. www.tdtc.ca/.

UN General Assembly. (1948). *Universal declaration of human rights*. https://www.un.org/en/documents/udhr/.

UN General Assembly. (1982). Principles of medical ethics relevant to the role of health personnel, particularly physicians, in the protection of prisoners and detainees against torture and other cruel, inhuman or degrading treatment or punishment. Resolution 37/194 of 18 December 1982. http://www.ohchr.org/EN/ProfessionalInterest/Pages/MedicalEthics.aspx.

Videto v. Kennedy, (1981). 125 D.L.R. (3d) 127 (Ontario Court of Appeal).

Vincent, N. (2012). Restoring responsibility: Promoting justice, therapy and reform through direct brain interventions. *Criminal Law and Philosophy*. doi:10.1007/s11572-012-9156-y.

Vrecko, S. (2010). Therapeutic justice in drug courts: Crime, punishment and societies of control. *Science as Culture, 18*(2), 217–232.

Warnick, J. E. (2007). Propranolol and its potential inhibition of positive post traumatic growth. *American Journal of Bioethics, 7*(9), 37–38.

Wolf, S. M., Paradise, J., & Caga-anan, C. (2008). The law of incidental findings in human subjects research: Establishing researchers' duties. *The Journal of Law, Medicine & Ethics, 36*(2), 361–383.

Section VI

History of Neuroscience and Neuroethics

History of Neuroscience and Neuroethics: Introduction

28

Frank W. Stahnisch

Contents

Abstract

The neurosciences have a long and fascinating history – one in which the bodily organ of the brain and its connection to theories about the soul and mind have triggered the interest of natural philosophers, physicians, researchers, as well as laypeople from ancient times to the modern period. As such, the history of the neurosciences – along with the recent history of neuroethics – incorporates wide perspectives from the history of philosophy and theology, the history of science and medicine, along with social, political, and cultural histories. This section will introduce the reader to several major topics in the complex history of ideas, theories, and neuroethical approaches to the structure and function of the brain and spinal cord. It first explores our modern knowledge about inflammatory and degenerative diseases of the brain, before eighteenth- to twentieth-century developments in basic neuroanatomy and neurophysiology are discussed. Mental health issues are also examined. There follows a description of the emergence of neurological surgery as a discipline towards the end of the nineteenth and beginning of the twentieth century, touching upon the particular new ethical

F.W. Stahnisch
Department of Community Health Sciences and Department of History, Hotchkiss Brain Institute/Institute for Public Health, The University of Calgary, Calgary, AB, Canada
e-mail: fwstahni@ucalgary.ca; frank.stahnisch@berkeley.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_22,

problems presented by operational manipulative and surgical approaches. The final two chapters explore the development of insulin and electroshock therapies, the history of new psycho- and neuropharmacological drugs since the 1950s, and the recent emergence of neuroimaging technologies – such as computer tomography (CT), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET). Their impact on the modern basic and clinical neurosciences will be described as well.

Introduction

The following section on the “history of neuroscience” for the “Handbook of Neuroethics” explores the long development of research and medico-scientific investigations into the brain and spinal cord. Such a historiographical overview can hardly be all encompassing, given the vastness of the research and clinical field, but some central landmark developments and transformations regarding the pursuit of knowledge in neuroscience are covered and explored with respect to their influence on the emergence of modern problems in the field of neuroethics. While trying to give here a sketch of some important foundations of brain anatomy and physiology, the emergence of clinical and surgical neurological sciences, particularly since the Early Modern Period, along with the innovative scientific accomplishments and methodologies emerging with the nineteenth century (see Coleman and Holmes 1988), this chapter attempts to shed light on the historical conditions that gave rise to the development of the new research field of “neuroscience” – as it appeared during the second half of the twentieth century. The socioeconomic, technological, and military contexts in which this special part of the modern life sciences itself became a “big science” are also highlighted.

Individual cultural and social backgrounds of the Western research endeavor in neurology, psychiatry, and later the neurosciences shall be investigated, giving meaningful attention to the way in which philosophical concepts, held by the public and the neuroscientists themselves, and practical and technological approaches to the brain and spinal cord influenced them. In mapping out the diverging historical episodes and various people and places in the long history of the neurosciences, the many side- and subdevelopments that occurred and their alignment with other research fields cannot be sufficiently covered in the five chapters of this section. Interested readers, however, who would like to know more about the changing cultural contexts of neurological and neuroscientific investigations, should refer to the more exhaustive accounts in the historiographical research literature (such as Brazier 1988; Finger 2000; Hagner 2000).

Discussion of the Individual Section Chapters

The individual chapters of this section focus on particular developments in the history of modern neuroscience and the related problems to which these have given

rise in the emerging field of neuroethics. Paul Foley's (University of Sydney, Australia) introductory chapter (▶ Chap. 29, "Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives"; ▶ Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"), for example, focuses on the study and treatment of movement disorders since the Early Modern Period as well as historical and ethical perspectives on Parkinson's disease since the nineteenth century. Foley gives a detailed review of both the diagnostic approaches to clinical phenomena related to diseases of the basal ganglia as well as their historical pharmacological therapies from the eighteenth century until the near present. He further shows that Parkinsonism affected a significant minority of elderly patients during much of human history, long before it had its modern definition – namely, as a progressive neurodegenerative disorder of the basal ganglia system of the brain and characterized through reduced dopamine levels in the corpus striatum structure of the diencephalon. Foley then presents suggestive evidence from various sources to this effect in the older literature, in patient files and from doctors' letters to their colleagues and patients. In his discussion, he dwells on precursor therapies such as ergot derivatives used since the Early Modern Period, other alkaloid therapies employed in the nineteenth century – beginning with atropine, hyoscine, scopolamine, and later the introduction of L-dopa – as well as various therapeutic approaches used during distinct phases. He also refers to the drastic infectious epidemic of encephalitis lethargica during the 1910s and 1920s.

The next chapter by Jean-Gael Barbara (Université Pierre et Marie Curie de Paris, France) (▶ Chap. 30, "History of Psychopharmacology: From Functional Restitution to Functional Enhancement") offers a short history of the development of psychopharmacology and early attempts to draw clearer diagnostic and therapeutic distinctions between diverse practices like drug administration to subjects of mental health institutions. Observations of the relationship between obvious somatic diseases and mentally ill patients through the longer course of the history of medicine are also made. Barbara then analyzes the development of the concept of psychopharmacology, as it appeared at the end of the eighteenth and the beginning of the nineteenth century, with more systematic and experimental investigations furthering pharmacological knowledge and artificial drug synthesis leading to closer examinations of the nature of the chemical substances and their influence on the treatment of mental diseases. This period marked an increased rationalization of psychopharmacological treatment options and understanding of underlying causal and physiological processes. Ethical discourses during the nineteenth century centered particularly around the problem of therapeutic nihilism – i.e., the refusal by contemporary physicians and pharmacists to use drugs, potions, and mixtures deemed harmful or at least ineffective – while issues of euthanasia were also discussed regarding instances of intractable pain, chronic and fatal diseases, as well as brain cancer and disseminated tuberculosis. In the third part of this chapter, Barbara addresses the rise of psychopharmacology as a discipline, which in turn gave rise to the increasing use of psychoactive drugs since the 1950s, as exemplified in the introduction of chlorpromazine, haloperidol, or reserpine. In each historical period discussed, multiple ethical questions arose regarding, for example, the specific risks

of the administered drugs, questions of addiction, and other adverse effects. Towards the end of the nineteenth century, clinical observations about mental diseases were more and more studied on the basis of systematic animal experimentation in physiology. Based on the innovative epistemology of experimental psychopharmacological research, complex clinical trial protocols and research techniques were developed, showing that this field of inquiry experienced intense progress while conversely raising multiple ethical problems which clinical neuroscience and psychiatry have been trying to solve ever since.

The third chapter by Delia Gavrus (McGill University, Montreal, Canada) (▶ Chap. 31, "Informed Consent and the History of Modern Neurosurgery") looks at human subject research in the history of neurosurgery and explores the long journey towards informed consent. Her chapter begins with an overview of the history of experimental approaches with human subjects in clinical and academic research settings from the late nineteenth to the mid-twentieth century, drawing specifically on foregoing work and publications by the American historians of science and medicine, Susan Lederer (1995) and Susan Reverby (2000). Gavrus then provides an introductory overview of the development of modern brain and spinal cord surgery, mentioning in that context new somatic therapies for mentally ill patients which gradually became available at the turn of the century. She proceeds to draw upon her prior work on brain and skull surgeries for criminals in the first decades of the twentieth century (Gavrus 2011), as these also gave rise to distinct ethical problems of their own. This article then explores the development of psychosurgery in the twentieth century (e.g., lobotomies, callosotomies, and nucleotomies) and raises the ethical issues that have been pointed out in the relevant scholarly research literature (see the work of Valenstein 1986; Pressmann 1998), as well as the changing doctor-patient relationship. Those changes influenced the ways in which neurologists, psychiatrists, and neurosurgeons informed their patients and test persons about the prognosis of their disease, the reach and depth of the treatment options, the nature of their proposed experimental and clinical treatment approaches, etc. This chapter ends on a note about the new psychosurgeries – deep brain stimulation, nerve tractomies, and destructive electrofrequency therapies.

Frank W. Stahnisch's (University of Calgary, Alberta, Canada) chapter, entitled "Non Restraint – Shock Therapies – Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry," (▶ Chap. 32, "Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry") discusses the early development of patient-centered treatments in nineteenth-century psychiatry and the backlash towards physician-oriented oppressive treatment and research styles in biological psychiatry at the beginning of the twentieth century. It is another aim of this chapter to trace some of the modern neuromanipulative approaches from deep brain stimulation back into their historical development and to compare them with more outmoded approaches regarding the electrophysiological alteration of the human cortex and deep brain structures. In the second part of this chapter, an overview on the development of modern deep brain stimulation methods and some of their major ethical problem fields is given. The neuroethical issue of external influences on human minds is

addressed from quite different interdisciplinary perspectives, such as from the basic scientific and clinical fields, from biomedical ethics, and from the philosophy of mind and brain science. This chapter adds another view to the discussion by putting forward a history of medicine and neuroscience perspective regarding the ethical problems involved, focusing on the issue of medical manipulation during the technological development of modern biological psychiatry. It is shown that, following the Second World War, electrophysiological stimulation approaches were developed, which began to crucially change the functional capacity of the human brain. By means of a comparative analysis, it is argued that many contemporary debates which question neuroethical applications (Glannon 2007) are problematic in significant respects. Of major concern, for instance, is the increasingly blurred conceptual boundary that is furnished by the complex relationships between clinical research, therapeutic intention, and physiological restitution.

In the final chapter, Fernando Vidal (Universitat Autònoma de Barcelona, Spain) (▶ Chap. 33, "Historical and Ethical Perspectives of Modern Neuroimaging") looks at a particular kind of innovative research technology – neuroimaging – which has become somewhat synonymous with the most cutting-edge research profiles and approaches of the modern clinical neurosciences. As Vidal intriguingly argues, neuroethics has been intimately connected with neuroimaging, both in its development and the definition of its tasks, and especially to the widespread application of functional brain imaging technologies such as PET and fMRI. Neuroimaging itself, in particular its uses, interpretation, communication, media presence, and public understanding, developed as one of neuroethics' primary subjects. Moreover, key neuroethical issues such as brain privacy or the conceptualization of blame, responsibility, and human personhood in general have largely gained from neuroimaging the form under which neuroethics deals with them. The employment of neuroimaging techniques to investigate phenomena usually associated with research in the humanities and human sciences brought those phenomena into the orbit of neurobiological explanation. Neuroethics emerged in the context of such technology-driven intellectual and professional developments. Thus, more than just being an important stimulus for the rise of neuroethics or a particular source of neuroethical challenges, the spread of functional neuroimaging can be considered a condition which reveals the sheer possibility of the field of neuroethics, as this chapter convincingly states. In return, neuroethics has come to support the claim that the neurosciences, particularly by way of functional neuroimaging, will revolutionize "traditional" ways of understanding the human condition. To the extent that such a claim is debatable, neuroethics might benefit these days from examining its special relationship to neuroimaging.

Conclusion

The individual chapters in this section collectively introduce major landmark developments in the long history of neuroscience that will be intriguing to anyone interested in modern issues of neuroethics. They provide historical and factual background information and insights into the social and cultural context in which

problematic new developments arose in modern basic and clinical neuroscience. Furthermore, major areas of antagonism between science and patient autonomy, physicians' freedom of decision, and external medical and pharmaceutical market forces, as well as underlying changes in the accompanying philosophies of the mind-brain relationship, are mapped and analyzed. From modern perspectives, these areas have generated particular ethical concern and ongoing discussions in the research field of neuroethics.

Cross-References

- ▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
- ▶ Historical and Ethical Perspectives of Modern Neuroimaging
- ▶ History of Psychopharmacology: From Functional Restitution to Functional Enhancement
- ▶ Informed Consent and the History of Modern Neurosurgery
- ▶ Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry
- ▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives

References

Brazier, M. A. B. (1988). *A history of neurophysiology in the 19th century*. New York: Raven.

Coleman, W., & Holmes, F. L. (Eds.). (1988). *The investigative enterprise: Experimental physiology in nineteenth-century medicine*. Berkeley, CA: University of California Press.

Finger, S. (2000). *Minds behind the brain: A history of the pioneers and their discoveries*. Oxford, UK: Oxford University Press.

Gavrus, D. (2011). Men of dreams and men of action: Neurologists, neurosurgeons and the performance of professional identity, 1925–1950. *Bulletin of the History of Medicine, 85*, 57–92.

Glannon, W. (Ed.). (2007). *Defining right and wrong in brain science: Essential readings in neuroethics*. Washington, DC: The Dana Press.

Hagner, M. (2000). *Homo cerebralis. Der Wandel vom Seelenorgan zum Gehirn*. Frankfurt am Main: Insel Press.

Lederer, S. (1995). *Subjected to science: Human experimentation in America before the Second World War*. Baltimore, MD: Johns Hopkins University Press.

Pressmann, J. (1998). *Last resort: Psychosurgery and the limits of medicine*. New York: Cambridge University Press.

Reverby, S. (Ed.). (2000). *Tuskegee truths: Rethinking the Tuskegee syphilis study*. Chapel Hill, NC: University of North Carolina Press.

Valenstein, E. S. (1986). *Great and desperate cures: The rise and decline of psychosurgery and other radical treatments for mental illness*. New York: Basic Books.

Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives

29

Paul Foley

Contents

Abstract

Therapeutic interventions employed in the management of extrapyramidal movement disorders have the potential to modify behavior, mood, and personality by virtue of the fact that the basal ganglia and dopaminergic neural systems affected by these disorders are also intimately involved in cognitive, emotional, and behavioral processes. The ethical implications of this situation, however, have received widespread attention only in recent years. Despite the sometimes negative changes in behavior that may ensue, the indubitable liberating motor benefits of dopamine replacement therapy employed to treat Parkinson's disease and the associated increase in personal freedom for the patient justify its employment. Neurotransplantation may in the future prove to be a useful approach that avoids such problems, but the current absence of a safe procedure militates against its use.

P. Foley
Unit for History and Philosophy of Science, University of Sydney, Sydney, NSW, Australia

Neuroscience Research Australia, Randwick, Sydney, NSW, Australia
e-mail: p.foley@neura.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_23,

Introduction

Voluntary motor activity in humans is controlled by two distinct neural systems: partially conscious control by the *pyramidal motor system* projecting from the motor cortex via the spinal cord to the muscles; and the unconscious maintenance of posture, as well as fine adjustment and coordination of motor activity, by the *extrapyramidal motor system*. In popular usage the term "movement disorders" elicits images of paralysis caused by disruption of the spinal innervation of muscles (pyramidal nerve cell disease). Neurologists, however, employ the term for conditions in which control of muscular tone or coordination is persistently impaired or abolished, despite the fact that the muscle involved and its direct innervation by the motor cortex are essentially healthy. These movement disorders result from dysfunction of the extrapyramidal motor system, the most commonly encountered extrapyramidal motor syndromes being Parkinson's disease (prevalence of 1 % in those over 60 years of age), essential tremor (5 % in those over 60 years of age), and tic disorders (up to 1 % of general population, more common in children and adolescents). Similar symptoms can also be unwanted consequences of prescribed medication, such as the tardive dyskinesias that develop in the course of antipsychotic therapy.

Hyperkinetic extrapyramidal motor symptoms range from small amplitude involuntary movements, such as tremor and myoclonus, to the extravagant motor excursions of athetosis, ballism, and chorea; it includes both rhythmic movements (tremor) as well as less ordered motions (chorea, tics), as well as complex actions that might appear voluntary were they not contextually so obviously inappropriate (the writhing, rolling movements of athetosis, the postural contortions of the dystonias). The persistent inability to remain still ("akathisia") is also a hyperkinetic extrapyramidal symptom, while *hypokinetic* extrapyramidal symptoms include the *slowing* (bradykinesia) or *absence* (akinesia, catatonia) of movement, as well as postural rigidity and instability.

The major nuclei of the extrapyramidal motor system are the *basal ganglia* – including the striatum (putamen and caudatus), pallidum, substantia nigra (SN), and subthalamic nucleus (STN) – connected in a series of complex circuits with each other and with other parts of the central nervous system, including the thalamus and cerebral cortex. These nuclei are, however, not solely concerned with movement, but are also involved in processes of psychological function (including motivation and reward), emotion, and automatic behaviors. The consequence is that an extrapyramidal disorder is typically anything but a purely motor disorder, and therapeutic interventions targeting motor symptoms may thus have unintended consequences in the mental and personality spheres.

The neurologist is bound by the same ethical considerations as other physicians: to reduce as far as possible the suffering of their patients, while minimizing any additional harm. The specifically neuroethical element enters these considerations when faculties particular to the CNS are involved, especially identity and personality. This chapter will therefore focus upon the impact of the best studied movement disorder, Parkinson's disease (PD), and its therapy upon these functions. PD was long regarded as an essentially motor disease but is now recognized to be the

"*quintessential neuropsychiatric disorder*", as it was described in a paper that documented the explosion of publications concerning cognitive, emotional, and sleep aspects of PD literature during the past two decades (Weintraub and Burn 2011). Neither PD nor its therapy can be discussed without considering the psychiatric consequences of each.

This overview will commence with a brief survey of how movement disorders were regarded in the nineteenth and early twentieth centuries, partly to provide background to later developments, but primarily to illustrate that the close linkage between disorders of mind and movement was recognized long before the neurologic basis for this nexus was clarified.

Extrapyramidal Motor Symptoms Prior to the 1920s

The extrapyramidal system was defined during the early 1920s, a successor to concepts of the "striatal" or "myostatic system" that had evolved between the late nineteenth and early twentieth centuries (overview: Lotmar 1926). Movement disorders had naturally occurred prior to the twentieth century, but were often interpreted as more mental than motor disturbances. More flamboyant forms, in particular, might lack obvious purpose, but these intricate motor excursions, often reminiscent of normal complex movements, nonetheless conveyed an impression of voluntary execution that suggested they might be the products of a *disordered* mind, but nonetheless of a *mind*. The frequency of movement disorders (tics, abnormal postures, bizarre motions) in mental patients was noted in an analysis of the latter nineteenth century English asylum casebooks, specifically in those inmates retrospectively classified as "schizophrenic" (Turner 1992). Phenomena that would today be regarded as "extrapyramidal" constituted an important part of the evolving definition of schizophrenia until the 1930s, and motor and autonomic symptoms now regarded as characteristic for extrapyramidal motor disease were seen as typical for mental disease (see Steck 1926, 1927). This was also evident in Kraepelin's procrustean accommodation of catatonia within his 1899 definition of dementia praecox, while Bleuler (1911) could recognize schizophrenia on the basis of gait alone. Even today the popular stereotype of the "madman" includes abnormal postures, gait, and speech among its most prominent features.

Movement disorders prompted the question of what else a person who had lost hegemony over their own body might be capable. This was especially true in the context of contemporary public and scientific fascination with somnambulism and hypnosis, and with emerging concepts of an unconscious aspect to the human mind. Those suffering movement disorders were thus not only objects of psychiatric scrutiny, but also forensic subjects, and their segregation was regarded as justified even where there was no evidence that an individual might pose an immediate threat to others.

It was, indeed, one of the most prominent investigators of clinical hypnosis who dominated the next phase in the interpretation of movement disorders. Jean-Martin Charcot was highly regarded for his contributions to delineating neurological

disorders (including "Parkinson's disease", a term he introduced), but is today perhaps more broadly remembered for his demonstrations of *la grande hystérie* at the Salpêtrière asylum in Paris. In the idealized form of this disorder, the patient presented a series of involuntary movements and changes in muscular tone that occurred spontaneously or in response to a specific stimulus, such as the touching of "sensitive spots" on their bodies. The dramatic writhing and arching of the woman's body, sometimes combined with attitudes of "ecstasy", constituted the very image of "madness", but the seeming ability of Charcot to initiate and terminate these states at will attracted both the interest of the general public and the criticism of colleagues who argued that Charcot had hypnotized his patients – if inadvertently, by force of his authority – and that the *grande hystérie* was a clichéd performance produced by impressionable patients to fulfil the expectations of their doctor (overviews in Trillat 1986; Goetz et al. 1995).

Important for the present discussion is that the Salpêtrière women did not exhibit a random assortment of bizarre motions, but rather a limited repertoire that included indubitably extrapyramidal symptoms. These "imitations" were sufficiently convincing that they could only be distinguished from the "genuine" neurologic symptoms by trained clinicians only on the basis of subtle differences and knowledge of the history of the patient. It is probable that, in some cases, they were not imitations, but expressions of organic basal ganglia disease, as in many respects they resembled the unusual motor symptoms of a neuropsychiatric disorder that emerged only a generation after Charcot's death: encephalitis lethargica. In this sense they were allied to the "neurosis" concept, applied to symptoms or disorders presumed to have an as yet undiscovered neurologic basis (such as PD). The allocation of a symptom to "hysteria" or "neurosis" was to a great extent dependent upon the context in which the symptom was presented.

Encephalitis Lethargica: The Omnibus Extrapyramidal Disorder

By the time of the First World War movement disorders could be regarded as neurologic or psychogenic, depending upon the clinician and the context in which the disorder was presented. There was also a growing interest at this time in the existence of neural pathways, distinct from the classical motor tracts, that were involved in maintenance of posture and the fine tuning of voluntary movement, spurred on in particular by Strümpell's concept of the "amyostatic symptom complex" (Strümpell 1915; see also Stertz 1921; Bostroem 1922) – essentially consonant with the more recent "extrapyramidal motor disorder" – and the landmark explorations of striatal neuroanatomy and neuropathology by the Vogts (1919) (overview: Lotmar 1926).

At precisely this moment, a new disorder challenged the separation of neurology and psychiatry and underscored the close relationships between the two halves of clinical neuroscience with respect to movement disorders. Two major forms of acute *encephalitis lethargica* (EL) were described during the peak of the 1918–1924 epidemic: a somnolent-ophthalmoplegic form dominated by symptoms

related to the impact of infection upon lower cranial nerves and brainstem sleep regulation centers, and a hyperkinetic form that additionally involved more anterior brainstem centers and, possibly, the striatum.

More pertinent to the present discussion were the *post-acute* symptoms of EL. Most of the 85 % who survived acute EL developed *post-encephalitic parkinsonism* (PEP), distinguished from PD primarily by the younger age of onset (15–35 years), and the severe accompanying autonomic symptoms (such as sialorrhea and seborrhea). The motor symptom palette of PEP was also broader than hitherto described for PD, including a variety of hyperkinesias, such as *primary restlessness* or *akathisia*, as well as a variety of *paroxysmal dyskinesias*, abrupt changes of muscular tone associated with clonic muscular contractions, the intensity of which generally abated during sleep, but which were exacerbated by emotional states or voluntary activity: *dysbasia lordotica* (thrusting forward of the pelvis); *torsion dystonias*; *oculogyric crises* (fixation of the eyes in an extreme upward or downward gaze); *cramps and tics*; and *respiratory abnormalities* (rapid, shallow breathing, barking, yawning). *Excitomotor syndromes*, involving more variable and extensive involuntary movements, and stereotypic behaviors, interpreted as reflecting the release of normally suppressed reflexes or motor patterns from higher control, were also typical for post-acute EL patients (see Zingerle 1936).

Every type of extrapyramidal symptom previously described in movement disorders, schizophrenia, and hysteria was thus also observed in EL (reviewed: Stern 1928; Guillain and Mollaret 1932). Such patients would previously have often been classified as psychiatric cases, but the epidemiology and neuropathology of EL undermined the view that such symptoms were psychogenic, even where their genesis and presentation were modified by emotional factors, and strongly suggested that even complicated, coordinated hyperkinesias, despite an apparently psychiatric or intentional nature, were indicative of extrapyramidal dysfunction (reviewed: Lotmar 1926; Stern 1928; Guillain and Mollaret 1932).

The symptomatology of post-acute EL also involved complex mental disturbances, most notably *akinesia* (lack of motor initiative) and *bradyphrenia* (slowed mental functioning). Careful interrogation of patients indicated that the general immobility of these patients was derived from disturbed interactions between the initiative and executive aspects of volition, a function hitherto regarded as primarily associated with the frontal lobe of the cerebral cortex (reviewed: Foley 2012). Some EL patients also presented symptoms normally encountered in schizophrenia, including catatonia and hallucinations, or classic hysteria, including the temporary amelioration of exacerbation of symptoms by suggestion, emotional factors, or pressure to certain parts of the body – not to mention the bodily contortions, arching of the back, and facial grimaces that characterized *la grande hystérie* (reviewed: Runge 1928; Steck 1931). These phenomena provoked debates about the roles of subcortical centers in "genuine" schizophrenia and hysteria, with the suspicion uttered by some that psychiatric disease was an expression of disordered subcortical function or of the loss of higher control of ontogenetically ancient subcortical behavioral responses (see, for instance, Schilder 1922; Reichardt 1928).

These suspicions were intensified by the behavioral syndromes exhibited by children and young adults who had suffered EL: prior to the development of motor symptoms, they presented a change in personality marked by impulsiveness, mischief, loss of social inhibitions, and attention deficits, combined with a sense of detachment from their own actions. That is, these children could steal, assault, lie, even murder and rape, despite an intact moral sense; after committing any of these actions, the child was passionately remorseful and convincingly insisted that they had been compelled to act as they did. This fissure between the conscious will and actual behavior could persist into adulthood, should the degree of motor incapacity allow its expression (reviewed: Thiele 1926; Fleck 1927b; see also Makowski 1983). Even many older PEP patients without major psychiatric problems retained an element of childishness or silliness, and were also markedly "clingy" and open to suggestion (see, for instance Fleck 1927a). It was frequently later noted that they long retained an exaggerated desire to please their doctors, and were often responsive to any new approach, regardless of how effective it proved to be in the longer term (Foley 2003).

The question naturally arose as to whether such persons were legally culpable for their actions, and European legislators generally concluded that they were not, although in England the application of such leniency needed to overcome an often unsympathetic judiciary on a case-by-case basis, not always with success. The nature of EL meant that those who developed asocial behavioral syndromes were not improved by punitive measures, as the "conscious self" was divorced from the undesirable activities of the individual. This ethical and legal conundrum was a source of ongoing discussion and consternation throughout the 1920s, and was ultimately resolved only by the unexplained disappearance of EL by the late 1930s (review of social aspects of EL: Neustadt 1932).

The phenomenology of EL demonstrated that emotivity, motivation, and motion are intricately interwoven in the subcortical brain, that topographic and functional lesions of the basal ganglia have consequences for the personality and its integration. This was of great practical concern during the 1920s: for instance, the largest EL outbreak, that of 1924 in the United Kingdom, aroused concerns that the generation following the one decimated by the Great War might itself fall prey to an infection that robbed young people of their mental autonomy.

EL was the epitome of extrapyramidal illness as a neuropsychiatric disorder, bundling as it did many of the problems encountered in less extravagant movement disorders. Many aspects of post-acute EL found their echo in phenomena that confronted clinicians decades later during the "modern era" of PD therapy.

Options in the Therapy of Parkinson's Disease

For many decades there were few options for the treatment of PD: belladonna alkaloid preparations (including atropine and scopolamine) were employed on the basis of their calmative and sedative properties as the mainstays of symptomatic management from the late nineteenth century until the early 1950s, when they were largely supplanted by synthetic anticholinergic and antihistaminergic preparations.

These agents provided only a modicum of relief by reducing resting tremor, and at the cost of dry mouth, loss of ocular accommodation, and gastrointestinal motility problems (see Foley 2003).

Dopamine Replacement Therapies

The 1960 discovery of the striatal dopamine deficit caused by SN degeneration, ultimately resulting in reduced thalamic activation of the cortex, remains the major watershed in the history of PD therapy. The importance of nigral degeneration in parkinsonism had been recognized by the 1930s, largely as a result of the neuropathological material provided by EL, but this initially had no impact upon models of PD. It was only in 1957 that Swedish pharmacologist Arvid Carlsson found that the listlessness induced by catecholamine depletion in rabbits could be reversed by restoration of CNS dopamine levels through administration of the catecholamine precursor L-DOPA; he concluded that a dopamine deficiency might also underlie parkinsonian akinesia, for which there was no treatment. In Vienna, pharmacologist Oleh Hornykiewicz discovered that dopamine levels were indeed markedly reduced in the basal ganglia of parkinsonian patients, leading to the initiation of a daring human experiment in 1961: neurologist Walter Birkmayer intravenously administered L-DOPA to a severely afflicted parkinsonian patient, with results that exceeded expectations, ultimately leading to the installation of L-DOPA as the "gold standard" medication for PD (reviewed: Foley 2003).

Although L-DOPA therapy was later hailed as the first "rational PD therapy", it should not be overlooked that these initial experiments provided part of the evidence that established its rationality, a bidirectional relationship between the principle and its proof that blurred the distinction between basic and therapeutic research.

The procedure involved is unlikely to be repeated today: a pharmacologist and neurologist, on the basis of laboratory findings partly raised by the pharmacologist, initiate an impromptu medical trial without consulting medical authorities, without a comprehensive research plan, and with a chemical supplied by its manufacturer for use in animals; there was no ethics committee from which to seek approval. The manufacturer was, indeed, suspicious that the dramatic effects achieved in this uncontrolled experiment had less to do with their powder than with the charisma of Birkmayer. As it transpired, this was not the case, but the international breakthrough for L-DOPA was not secured until neurologist George Cotzias achieved positive results with oral DOPA in the United States. His work was, ironically, based on a false hypothesis (he aimed to increase melanin levels) and undertaken with rather more recklessness: he increased the dose until the desired effect was achieved, but also a great deal of unpleasantness:

> Sure, the patients are suffering, but hell, what they're going through ain't nothing compared to what I went through when I was a sergeant in the Royal Greek Army. When the patients reach the degree of suffering that I experienced as a sergeant in the Greek Army, then we'll stop the study. But right now there's too much at stake (cited by Foley 2003, p. 490).

The single-minded personalities of Birkmayer, who charmed his patients, and Cotzias, who dominated them, played decisive roles in establishing L-DOPA therapy. But both were representatives of the model of drug discovery and introduction that was already waning with the waxing dominance of pharmaceutical firms since 1945, but also with respect to the ethical framework for such research. The requirements for pharmaceutical trials in the United States stipulated by the Kefauver-Harris Amendment to the Food, Drug, and Cosmetic Act in 1962, for instance, included prior notification of the names of investigators and their qualifications, the structure and justification of the planned investigation, the obtaining of written informed consent from all patients and other subjects involved in trials. This prompted the New York neurologist Lewis Doshay, who had spearheaded the introduction of the synthetic anticholinergic agent Artane (benzhexol) into PD therapy in the early 1950s, to berate these changes as counterproductive and unnecessary:

> [in the 1950s] we had none of the current problems of obtaining written consent from the patients and their families, in order to test a new drug. Nor were we required to explain to them that it was a new and unknown drug containing many potentially dangerous side reactions. Our patients ... trusted us and were entirely confident that we would not give them anything to harm them. As a matter of fact, in the course of 30 years we had tested over 300 new drugs without a single instance of a serious toxic effect in thousands of patients (Doshay 1965).

In his opinion, a "*competent and careful investigator*" could achieve evaluations which were at least as informative as objective measuring devices by careful long-term, large-scale clinical investigations. Doshay was dismayed by the regulations which in his opinion had begun to hamper research, and felt that *"freedom of investigation no longer exists and the patients wait in vain for new and better remedies"* (Doshay 1965).

Whatever retrospective ethical doubts might be entertained, the unprecedented success of L-DOPA, improving not just the mobility of the sufferer but also allowing them a longer, more active life outside institutional care, subsequently led to the use of other agents that enhance CNS dopaminergic transmission (peripheral decarboxylase inhibitors, monoamine oxidase inhibitors (MAOI), dopamine receptor agonists), and *dopamine replacement therapy* (DRT) is today the dominant paradigm in the pharmacological therapy of PD (review of therapeutic options: Jankovic and Poewe 2012).

Nevertheless, L-DOPA does not cure PD – it neither arrests nor reverses the demise of the SN – and is itself associated with significant adverse events, the most significant being the development within a few years of abnormal involuntary movements (*tardive dyskinesias*), as well as of *on-off fluctuations* in the effectiveness of the agent. It was also suggested during the 1980s that L-DOPA itself might be cytotoxic (by increasing local oxidative stress), and thereby contribute to the progression of the disease, although this view is no longer so widely held. Given the undoubted benefits of L-DOPA therapy, and as the duration of PD prior to initiation of L-DOPA therapy is not associated with significant differences in response, it therefore seems advisable to postpone initiation of its use for as long as possible by employing alternative DRT measures, despite their being less effective and more expensive options, and not without their own problems.

In particular, dopamine receptor agonist medication (but not MAOI therapy alone) has been associated with *impulse control disorders* in 15–20 % of treated patients (Carter et al. 2011; also see reviews listed below). The most frequently reported forms include *increased sexual activity*, including proscribed practices, such as pedophilia; *problem gambling* or *shopping*; *compulsive eating*; and *excessive devotion to hobbies*. Increased tobacco, alcohol, and drug use, curiously, are rare. The most frequently encountered problem is pathologic gambling, the most popular gambling form being poker (slot) machines, although internet gambling is expected to assume a more prominent role in the future; younger patients and males are at particular risk (incidence as high as 9 %, compared with 0.5 % in the general population), consistent with a normally higher level of risk-taking and addictive behavior in these groups, and of especial significance when employing dopamine receptor agonists in early onset PD or restless legs syndrome. It is presumed that genetic, personality, and personal history factors (including a family history of addiction disorders) play at least a predisposing role, as do social factors, as indicated by the fact that gambling addiction is not as frequently a problem in societies where gambling is illegal (Appel-Cresswell and Stoessl 2011; Raja and Bentivoglio 2012).

As early as the late 1960s, it had been recognized that L-DOPA therapy was associated with increased libido in some patients (not always regarded as an adverse drug effect), but it is more commonly associated with *punding*, the persistent execution of particular stereotyped behaviors (such as cleaning, hoarding, grooming, or sorting), often related to previous work or hobbies; the term was coined by Swedish psychostimulant (amphetamine and cocaine) users to describe their own pointless perseverations. Punding is less disturbing for the patient than an impulse control disorder, as it does not involve invasive thoughts, but irritation may ensue should their stereotypic behavior be blocked or criticized. *Dopamine dysregulation syndrome* (DDS; also "hedonistic homeostatic dysregulation"), or excessive use of L-DOPA (c. 4 % of treated patients), resembles drug addiction in that patients persist with high dosing despite the resulting severe dyskinesias, and autonomic and mental withdrawal symptoms may accompany discontinuation or dose reduction.

DRT-treated PD patients may also present combinations of these behavioral abnormalities, and they are generally accompanied by more general psychiatric symptoms – particularly depression, anxiety, apathy, irritability, and executive impairment – symptoms that five in six non-demented Parkinson's disease patients, however, present within 7 years of diagnosis (review: Bonnet et al. 2012).

There is no doubt that these behaviors are triggered by DRT: the typical personality profile of PD patients is one of reserve, risk avoidance, and low addiction potential; further, the problem behavior typically normalizes following cessation of treatment. This does not mean that the patient is a priori not responsible for untoward behaviors; it might be argued, as did some authors in the case of EL behavioral syndromes, that the medication merely uncovered or released a pre-existing proclivity in the personality of the patient. For example, an English school headmaster was found to possess several thousand items of child pornography on his computer, only one of which had been downloaded prior to DRT: the question

arose as to whether this item was evidence of pedophilic tendencies prior to treatment (Carter et al. 2011). Another PD patient insisted that his fascination with anal intercourse during DRT was consistent with previously unexplored desires; after a change in medication, however, he not only lost his newfound interest, but was also ashamed that he had ever experienced such feelings (Carter et al. 2011).

DRT aims to simulate restoration of transmission in the nigrostriatal pathway, but can also be expected to have an impact upon a second major dopaminergic system, the *mesocortical-mesolimbic system*, excessive activity in which was implicated in the early 1960s in the etiology of psychosis, and which is now known to be involved in motivation, reward, and learning mechanisms. Although the concept of dopamine as a "reward" or "happiness chemical" is overly simplistic, this system is undoubtedly important for proper emotional performance as well as for goal-directed behavior, and genetic variations in dopamine-mediated transmission (release, reception, inactivation, re-uptake) are believed to be involved in predisposition to a variety of behavioral and emotional conditions, including impulsive-compulsive behaviors and depression. Further, dopaminergic neurons in the prefrontal cortex are involved via their regulation of information flow between CNS regions in memory and attention processes (Foley 2009). The ability to learn from positive outcomes is reported to be impaired in PD, contributing to their inertia, while DRT reduces the capacity for learning from negative experiences, so that correction of harmful behaviors and decision-making is impaired. Further, increased ventral striatal dopamine levels are associated with increased novelty- and reward -seeking (the left striatum is possibly more important in this respect, while the right is associated with harm minimization); as dopamine levels are not reduced in this region in PD, DRT may produce a relative "overdosage." Functional changes in the orbitofrontal cortex may be involved in the genesis of such behaviors. Nevertheless, it should be borne in mind that the neurobiology of impulse control disorders in general remains unsettled, so that the mechanisms underlying such behaviors in PD cannot be regarded as completely elucidated (Bodi et al. 2009; Cilia and van Eimeren 2011; Voon et al. 2011; Raja and Bentivoglio 2012).

Impulse control disorders and related conditions encountered during DRT can be catastrophic for the patient, with financial ruin, legal proceedings, and family breakdown not infrequent consequences of problem covert gambling or altered sexual behavior. They differ from the "bad behavior" of EL syndromes in that the PD patient generally identifies with their altered behavior during DRT, even should it be alien to their personality before and after therapy; that is, there is no perception that the behaviors are not "owned" by the patient. It is nonetheless difficult to clearly determine whether altered behaviors should be attributed to a change in personality, in volitional processes, or in the ability to suppress inappropriate impulses. In other words, the degree to which DRT medication occludes decision-making processes in general, thereby reducing the agency or autonomy of the patient, is unclear; that is, whether it not only increases the inclination to a particular behavior, but also reduces the ability to resist temptation. It is currently

thought that increased impulsiveness in DRT patients is combined with a lowered ability to censor their own behavior, so that recognition of reduced social and legal and competence is given. This might be compared with the impaired responsibility of a person induced by alcohol or other recreational drug that reduces the capacity for disciplined decision-making, except that the PD patient is taking prescribed medication for a serious medical condition, and does not enjoy the same degree of freedom to choose abstinence as a solution to their problem.

In practice this poses no major ethical problem, as these untoward effects can usually be resolved by switching the patient to another medication: further, the effectiveness of this approach confirms that the patient had not *chosen* to offend. Issues of "change of identity" have not been raised in connection with DRT (in contrast to DBS; see ▶ Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"), as it is clear that DRT only temporarily and reversibly alters aspects of personality in a manner comparable with other intoxicants, not the core personality of the patient.

As it is currently not possible to predict which patients will respond in this manner, it is imperative that DRT recipients' behavior be carefully monitored both by their family and their physicians, particularly as titration of dosage or switching to an alternative agonist alleviates the situation in most patients. As insight and decision-making are compromised, screening before and during medication is important. It may also be necessary to institute precautionary measures, such as transferring legal control of financial resources to guardians prior to initiation of therapy, particularly as patients with gambling problems are adept at concealing their often not inconsiderable losses. Whether medication control is required in DDS cases where excessive dosage does not result in problematic behaviors is questionable: if the patient accepts motor dyskinesias as the price of their abuse, restricting their freedom in order to preserve their capacity to make rational decisions is in itself problematic, particularly as "harm minimization" strategies, rather than abstinence, are currently popular with respect to illicit recreational drugs.

This issue of the responsibility of pharmaceutical manufacturers in this regard requires little philosophical analysis, as the manufacturer of *any* pharmaceutical agent is clearly obliged to disclose all *known* adverse effects of such agents prior to supply, and impulse control disorders certainly fall into this category. The role of dopaminergic systems in addiction and reward mechanisms were investigated as early as the 1970s, but it was not until the early 1990s that the major role of mesolimbic dopaminergic circuits in these areas was firmly established (review of current state of knowledge: Volkow et al. 2012). From this point it could be expected that any agent that modified dopaminergic tone might have an impact upon behavior. Courts have been petitioned by patient groups in several countries to decide whether manufacturers were aware of the behavioral risks associated with dopamine receptor agonists prior to their commercial release; this, however, is a forensic, not an ethical issue.

Reviews of impulse control disorders and DDS associated with DRT: O'Sullivan et al. 2009*; Voon et al.* 2011*; Raja and Bentivoglio* 2012*; Weintraub and Nirenberg* 2013.

Neurotransplantation

Neurotransplantation (neural grafting) aims to replace nigrostriatal cells lost to degeneration, in the hope that appropriate connections with existing nerve cells ensue, re-establishing normal striatal function. Depending upon the degree to which the graft restored physiological dopamine release, it would allow reduction or even abolition of DRT, together with its negative correlates.

The idea was not entirely new when it was introduced in the late 1980s: a bovine striatum extract for the treatment of PD had been marketed in Germany in the late 1920s under the trade name "Striaphorin", taken as three tablets per day. Its chief proponent, Berlin neurologist Heinrich Rosin, reasoned that, if the material survived digestion, *"it is possible that degenerating centres in the central nervous system can still be rescued or that such centres that are liable to degeneration can be preserved by administration of healthy nervous substance with the same function"* (Rosin 1930). Despite some encouraging results, the expensive therapy was never adopted on a widespread basis, and any benefits can presumably be attributed more to the impressionability of PD patients than to its specific qualities. It may nevertheless be validly viewed as the conceptual precursor of more recent neurorestorative approaches.

The potential of neuronal grafts for alleviating the dopamine deficit had been explored during the 1970s in an animal model of parkinsonism, the 6-OHDA-lesioned rat, and the results were promising: the grafts survived and flourished, restoring normal striatal dopamine levels; normal electrophysiological firing patterns were established; and, most importantly, the motor deficits of the model were significantly ameliorated. The first clinical trials in PD patients followed in the early 1980s, employing cells from the patient's own catecholamine-synthesizing adrenal medulla tissue, but the results were unimpressive. This led to consideration of the use of human fetal nigral cells, not expectedly triggering an ethics debate. The Swedish Society of Medicine concluded that the strategy was conscionable if the decision to abort the fetus had been independent of the intended use of its cells, if the woman carrying the fetus provided consent, and if the entire procedure was conducted on a noncommercial basis. These guidelines later informed the first international guidelines for the ethical use of fetal tissue in transplantation (reviewed: Dunnett 2010).

Implantation of fetal ventral mesencephalic tissue (8–10 weeks postconception) into the putamen, the striatal component with the greatest dopamine loss in PD, proved to be quite successful: the grafts survived, striatal dopamine levels were increased, and there was a marked improvement in L-DOPA "on" time. These benefits were maintained for at least ten years after surgery. Further equally encouraging open-label studies followed, the findings of two double-blind, placebo-controlled studies instituted by the National Institutes of Health (USA), however, cast unexpectedly serious doubts on the future of neurotransplantation for PD: not only were the clinical benefits insignificant and graft survival poor, several patients developed severe "runaway graft-induced dyskinesias", initially ascribed to uncontrolled proliferation of the transplanted cells. Press reports garishly

declared that the therapy once hailed as a "miracle" had instead proved to be a nightmare. The relative success of deep brain stimulation at about the same time also undermined confidence in the risk-benefit balance of neurografting.

Transplantation trials have since been suspended until the problems brought to light by the NIH trials have been resolved. Research in this regard has included a shift in focus to deriving dopamine-producing cells from human embryonic stem cells (from embryos produced in the laboratory in the context of in vitro fertilization), thereby securing a better defined and more consistent cell source; the possibility of restoring stem cell-like properties to adult cells ("induced pluripotent stem cells") has also raised hopes that ethical reservations regarding material derived from human fetuses or embryos can be circumvented (Moreira and Palladino 2005; Brundin et al. 2010; Lindvall and Björklund 2011; Politis and Lindvall 2012).

Neurotransplantation can certainly not be equated with the psychosurgery of the 1950s, where large regions were denervated or removed without respect for the far-reaching consequences of such adventurism; modern neurografting is a cautious attempt to compensate pathological nerve loss in the brain (or spinal cord, where it does not incur the same rancor). Proponents indeed argue that local extracellular signalling directs differentiation and connection of the grafted cells, so that the process can be regarded as one of self-repair, whereby the applied allograft supplies only the raw materials.

If an effective, safe methodology can be developed, the next practical issue concerns the selection of patients for the new trials. Advanced disease patients experiencing unsatisfactory treatment response may seem the most appropriate candidates, as they are presumed to have less to lose than less afflicted patients. In the case of a therapy predicated upon self-restorative capacities, however, younger subjects would be preferable, especially as significant restoration may not be possible where the disorder is too far advanced.

The patient's attitude to risk would also be critical in assessing their suitability, given the potential for irreversible unforeseen complications of neural transplantation; full disclosure of the problems encountered in the NIH trials would be among the information supplied. Although general public uneasiness with respect to neurosurgery can be expected to dampen exaggerated enthusiasm for partaking in clinical trials of neurotransplantation, the patient may nonetheless be swayed by the neurologist's support or even enthusiasm for a new strategy, so that it is essential that explanation of the extent of the expected benefits and the irreversibility of the procedure, as well as the recording of consent, involve not just their neurologist but also other health specialists, dedicated counsellors, and the patient's family. This is particularly important as the motor benefit of neurotransplantation was not superior to standard L-DOPA therapy: *"Transplants improved objective signs of Parkinson's disease to the best effects of L-DOPA seen preoperatively. ... In subjects in whom transplants replaced the need for L-DOPA, the implants replicated the preoperative effects of L-DOPA, including dyskinesias in susceptible patients"* (Freed et al. 2011). That is, the major advantage lies in the convenience of reducing the need for exogenous L-DOPA, although total withdrawal was not possible in all cases.

Despite the conceptual support for the approach from animal studies and pre-2003 experience of mesencephalic grafting, the implantation of nerve cells into the human brain constitutes "human experimentation" to an even greater extent than deep brain stimulation therapy of PD, and this must also be explicitly explained to the patient and their family. Master et al. (2007) discussed the potential benefits to society of stem cell implants with regard to the gain in knowledge of the CNS, but the epistemic dividend of the technology should be handled cautiously; the medical benefit for the individual must remain paramount, particularly as the patient assumes the entire risk of the intervention (hemorrhage or infection during surgery; undesirable immunologic responses; inappropriate behavior of implanted cells, including tumor formation and cell migration), while their medical condition may not necessarily improve significantly even if surgery is successful.

Some authors have argued that neurotransplantation is little more than an attempted rehabilitation of the discredited practice of psychosurgery (see, for instance, Linke 1993; Moreira and Palladino 2005). Their stance is based upon their equation of the biological term "self" and the philosophical concept of "identity", arguing that the introduction of foreign tissue into the brain cannot be regarded as involving "self-repair", so that it necessarily alters the brain and consequently the identity associated with that brain (cf. detailed discussion of relevant issues by Frankfurt psychiatrist and philosopher Georg Northoff 1996, 2004). The current view of the dynamic relationship between CNS structure and function renders it unlikely that insertion of even a few thousand dopamine-releasing cells into the striatum would significantly affect a patient's sense of identity, as the cells are not additional to the normal complement, but essentially replace neurons lost at a site not associated with mental functions – which loss itself, incidentally, did not appear to directly impact upon the patient's sense of identity. Optimism is also encouraged by experience with striatal neurotransplantation to date, both by the absence of major psychiatric effects and by the observation that the positive and negative impacts of successful grafting are comparable in quality and quantity with those of normal L-DOPA therapy. This is not to deny that neurotransplantation in certain areas of the brain *could* modify personality or mental functioning, and thereby a patient's sense of identity. This, however, applies to any therapy, pharmacological, surgical, or psychological, that modulates dopaminergic transmission. Neurotransplantation might well be preferred to L-DOPA therapy precisely because it modifies dopaminergic parameters only at the site of insertion, and not in mesolimbic or cortical regions, as is the case for L-DOPA therapy.

A final ethical issue concerns the sham surgery involved in the "placebo-controlled trials" of transplantation (and also of newer strategies, such as gene implantation therapy). Sham surgery, including adjunct anesthetic, antibiotic, and immunosuppressive medication, clearly entails a greater risk to a patient than a placebo pill, and it has been argued that pre- and postsurgery comparisons of predefined parameters would be more appropriate. A particular problem associated with controlled surgical trials is that, despite the costs and inconvenience incurred, it must be possible for the subject to withdraw consent for

participation in the treatment or placebo arms of such a trial, even after surgery has commenced (brain surgery is normally conducted while the subject is conscious) (Rabins et al. 2009; further discussion: Dekkers and Boer 2001; Albin 2002; Frank et al. 2008; Brundin et al. 2010; Galpern et al. 2012; Swift and Huxtable 2013).

Reviews of neural transplantation in PD: Brundin et al. 2010*; Mauron and Hurst* 2010*; Lindvall and Björklund* 2011*; Politis and Lindvall* 2012.

Deep Brain Stimulation

Considerable excitement has been generated in the past two decades by the management of parkinsonian symptoms by means of deep brain stimulation (DBS), usually directed at the subthalamic nucleus. This therapeutic approach employs electrical modification of basal ganglia circuit activity rather than direct modulation of dopaminergic transmission, although the latter is potentially involved in its mechanism of action. Concerns have been expressed that the alteration of mood often encountered as a by-product of this intervention may represent an inappropriate intrusion into or alteration of personality, and may also alter cognitive performance. The neuroethical issues associated with DBS therapy for PD will be discussed in detail in ► Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"; here it suffices to note that the psychiatric effects of DBS, in any case a last choice therapeutic strategy for PD, are generally not a cause of lasting concern for either the patients or their families, although adjustment to the improved mobility may be required of both parties. Ethical issues regarding the use of DBS to alter behavior or mood in non-neurological disease cannot be directly extrapolated to its specific application to the relief of movement disorders.

Overview of Neuroethics of Therapy for Parkinson's Disease

Four principles have established themselves in medical ethics as useful guides: patient autonomy, non-maleficence (do the patient no harm), beneficence, and justice or fairness (Beauchamp and Childress 1994).

Autonomy

Two major issues are included in this category, the first being that of the ability of the patient to provide *informed consent* for a medical procedure. PD patients have been reported since the 1920s to be more open to or even demanding for novel therapies, and to be more responsive to placebo treatments (including sham surgery), even on objective measures (such as neuroimaging), and this should be considered when discussing consent. Dementia is a feature of latter stage PD, but any therapy is typically initiated long before this point; it is, in any case, an exclusion criterion for the initiation of neurotransplantation and DBT therapies.

The PD patient can be expected to feel anxiety and perhaps desperation when initially diagnosed, and this anxiety will be revived and intensified with each perceived therapy failure. Further, their incomplete understanding of the nature of their disease impedes the ability of the patient to make a genuinely informed decision, but this applies to most interactions between layperson and physician.

Press reports of the "miracle" effects achieved by particular therapies complicate informed consent to an even greater extent, so that the possibility and nature of side effects must be carefully discussed, and the patient's immediate family involved in such discussions. The nature of the surgical techniques employed in PD also means that treatment will be provided by a limited number of centers, presumably including particularly confident advocates of these procedures. This contrasts with the situation for new medications; these can be trialled in a large number of patients in different settings, rapidly providing a considerable material concerning positive and adverse effects.

The second issue is whether *patient autonomy is compromised by therapy*. Patient autonomy is in general enhanced by therapy as a result of the improvement in motor freedom and enhanced mental initiative, but DRT therapies can elicit impulse control disorders in which the volitional freedom of the patient is challenged; this effect, which may reflect relief of disease-related behavioral inhibition, can be countermanded by modification of therapy.

Non-Maleficence

The *selection of therapy* must consider the nature of the therapy, the potential side-effects, and the current state of the patient. Surgical therapies are more invasive and, for the foreseeable future, more experimental than DRT therapies, so that caution must be exercised with respect to the physical, cognitive, and emotional status of the patient, including their attitudes to surgery and risk.

While PD is a grave disease, the risks associated with *not* implementing novel therapies are not so great that they justify unnecessarily risky alternatives (Kimmelman et al. 2009). As with other CNS-based disorders, the results of animal models can only be extrapolated to humans with reservations, so that these must be well-designed investigations capable of providing secure data, and their results interpreted conservatively. Human trials should be carefully designed to maximize both safety and their ability to yield reliable results that can guide further investigation, compensating the risks inherent in such trials. Finally, while such trials contribute to expanding knowledge of CNS function, such as the impact of certain interventions upon CNS dopaminergic parameters, clinical benefit for the patient remains the most important guiding principle when evaluating the risk-benefit relationship of an intervention.

Beneficence

This aspect is assured by implementing only those therapeutic strategies that both have a scientific basis – that is, there is a rational basis for the approach, in most cases derived from animal models of the disease in question – and controlled, double-blind trials have confirmed the clinical efficacy of the strategy (class I evidence).

The latter condition can naturally not be met by experimental treatment strategies, such as neurotransplantation, so that it is all the more important that clinical trials build upon a secure scientific basis, including extensive animal investigations. Clear consensus decision strategies with respect to treatment options have been established in both United States and Europe that also serve as guides elsewhere, and a conservative approach, commencing with the familiar L-DOPA and only proceeding to more invasive techniques only where absolutely necessary, minimizes the risk to patients (Fins 2009; Bell et al. 2011).

Justice

The major factor in this respect is the question of whether the benefit afforded by therapy justifies the costs involved. No physician or patient would contemplate a return to the pre-L-DOPA era, when PD often resulted in permanent physical incapacity and institutionalization. Conservative pharmacological therapy can thus be justified on this basis without further discussion, the only limitation being the financial capacity of the patient and their local health economy. Given their current developmental status, the question of whether the costs of surgical therapy justify their use in place of normal pharmacological therapy cannot yet be answered.

Ethical Challenges of the Early Identification of Parkinson's Disease Candidates

Genetic and brain imaging techniques provide researchers with the potential to discover nascent disease before clinical symptoms develop. Around 10 % of PD cases are genetically determined, and genetic counselling of family members is appropriate with regard to life and family planning, particularly as hetero-allelic inheritance may also be associated with dopaminergic dysfunction, while those with certain dominantly inherited mutations may remain healthy into old age. Nevertheless, the limited value of genetic testing in predicting outcome means that it is more important in research contexts than in the clinical setting (Appel-Cresswell and Stoessl 2011).

Brain imaging techniques can detect structural or functional abnormalities that may identify nascent disease in asymptomatic persons (family members of PD patients, or probands in a research setting). As with detection of relevant genetic mutations, the conundrum arises as to whether "healthy" persons should be informed about the detection of "abnormalities", even though it cannot be certain that the identified phenomena will lead to clinical disease, or may even prove, after further research, to be within the limits of normal physiological variation. As noted by Appel-Cresswell and Stoessl (2011), identifying a functional dopaminergic abnormality cannot be equated with discovering a tumor, as the significance of the former will be a matter of interpretation based on a complex of assumptions and technical features. This scenario must be thoroughly discussed before initiation of an investigation, and then discussed with those to be tested before seeking their consent for participation. The issue is particularly difficult because there are no

measures that could be recommended to the proband in response to such a result that would halt the course of any future disease (contrasting, for instance, with the discovery of a tumor or vascular issue for which medical treatment might be available), so that they pass from a situation of undisturbed ignorance to one of disconcerted ambiguity.

Unless the proband has indicated in advance that they would definitely wish to be advised of any abnormalities discovered – relatives of PD sufferers, or biomedical professionals, might participate in such studies, for example, specifically to ascertain their status – it is probably advisable to not disclose specific research findings. This approach, however, is still encumbered by the difficulty that many participants, knowing the purpose of the investigation in which they are participating, will presume that they will be told of *any* abnormalities, regardless of prior discussions.

Finally, the current state of knowledge militates against employing brain imaging techniques for the solution of legal questions regarding the etiology of parkinsonian conditions: the intentions of exploratory research remain separate from those of the application of its results until those findings have been substantiated.

Concluding Remarks

The yardstick of ethical behavior with respect to movement disorders, as for all medicine, is that the physician applies their available energy and resources to reducing the suffering of their patient, and to do so, as far as possible, without introducing new harm into their lives. Neuroethics can thereby be regarded as simply the application of general medical ethics to disorders and functions of the brain. The specifically neuroethical element of such discussions emerges where features and activities that are particular to the brain are involved: cognition, mood and emotion, and particularly identity and personality. Magic bullets, however, are rare in the clinical neurosciences, so that effective therapy is accompanied by "undesirable side-effects", and the question for both doctor and patient concerns the degree to which the latter is willing to accept these adverse consequences in order to alleviate existing discomfort. What appears to be a simple issue of the balance between positive and negative effects proves, in practice, to be somewhat more complicated, but many apparent ethical conundrums can be circumvented by retaining a pragmatic view of the patient as a real person with real medical problems that require solutions, and not to burden clinical decision-making with excessive application of metaphysical analysis of situations, when asking the patient how they felt would be more helpful.

Two principles can serve as rough guides: that therapy should be no more invasive than necessary and as reversible as possible; and that it should always be based upon the latest available understanding of both CNS function and the mechanisms of the strategy employed. No medical intervention is natural: strictly natural would be to allow PD patients to be reduced to immobile cripples, as was the

prospect for most prior to the advent of L-DOPA therapy. For the foreseeable future, there will be large gaps in our knowledge of the both PD and its neural substrate, as well as of the mode of action underlying the therapies applied. But neither these gaps nor the risk of side effects of any type justify outright prohibition of these interventions, as this would amount to therapeutic nihilism and deprive PD sufferers of any hope of relief: it is not to be expected that all the unknowns involved in PD and its treatment will be resolved in the foreseeable future. It would be as paternalistic to withhold promising advances in therapy because of residual uncertainty as it once was to run roughshod over individual patient concerns. One should also not underestimate the willingness of PD patients to participate in trials of new therapies, even in this air of uncertainty; the fact that they have PD means that they have already tolerated a great deal, and many are willing to take chances to advance both their own welfare and their quest for personal autonomy, as well as that of their fellow sufferers.

Cross-References

▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Ethical Implications of Cell and Gene Therapy
▶ Ethics of Sham Surgery in Clinical Trials for Neurologic Disease
▶ Impact of Brain Interventions on Personal Identity

References

Albin, R. L. (2002). Sham surgery controls: Intracerebral grafting of fetal tissue for Parkinson's disease and proposed criteria for use of sham surgery controls. *Journal of Medical Ethics, 28*, 322–325.

Appel-Cresswell, S., & Stoessl, A. J. (2011). Ethical issues in the management of Parkinson's disease. In J. Illes & B. J. Sahakian (Eds.), *The Oxford handbook of neuroethics* (pp. 575–600). Oxford: Oxford University Press.

Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. New York: Oxford University Press.

Bell, E., Maxwell, B., McAndrews, M. P., Sadikot, A. F., & Racine, E. (2011). A review of social and relational aspects of deep brain stimulation in Parkinson's disease informed by healthcare provider experiences. *Parkinson's Disease, 2011*, ID 871874 (8pp.).

Bleuler, E. (1911). *Dementia praecox, oder Gruppe der Schizophrenien*. Leipzig, Wien: Franz Deuticke.

Bodi, N., Keri, S., Nagy, H., Moustafa, A., Daw, N., Dibo, G., Takats, A., Bereczki, D., & Gluck, M. A. (2009). Reward-learning and the novelty-seeking personality: A between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. *Brain, 132*, 2385–2395.

Bonnet, A. M., Jutras, M. F., Czernecki, V., Corvol, J. C., & Vidailhet, M. (2012). Nonmotor symptoms in Parkinson's disease in 2012: Relevant clinical aspects. *Parkinson's Disease, 2012*, 198316 (15pp.).

Bostroem, A. (1922). Der amyostatische Symptomenkomplex. Klinische Untersuchungen unter Berücksichtigung allgemein-pathologischer Fragen. Berlin: Julius Springer.

Brundin, P., Barker, R. A., & Parmar, M. (2010). Neural grafting in Parkinson's disease: Problems and possibilities. *Progress in Brain Research, 184*, 265–294.

Carter, A., Ambermoon, P., & Hall, W. D. (2011). Drug-induced impulse control disorders: A prospectus for neuroethical analysis. *Neuroethics, 4*, 91–102.

Cilia, R., & van Eimeren, T. (2011). Impulse control disorders in Parkinson's disease: Seeking a roadmap toward a better understanding. *Brain Structure & Function, 216*, 289–299.

Dekkers, W., & Boer, G. (2001). Sham neurosurgery in patients with Parkinson's disease: Is it morally acceptable? *Journal of Medical Ethics, 27*, 151–156.

Doshay, L. J. (1965). Problems in Parkinson's disease confronting the profession and the community. In A. Barbeau, L. J. Doshay, & E. A. Spiegel (Eds.), *Parkinson's disease. Trends in research and treatment* (pp. 1–5). New York: Grune & Stratton.

Dunnett, S. B. (2010). Neural transplantation. In S. Finger, F. Bolle, & K. L. Tyler (Eds.), *History of neurology* (Handbook of clinical neurology, Vol. 3, pp. 885–912). Elsevier: Edinburgh. Vol. 95.

Fins, J. J. (2009). Deep brain stimulation, deontology and duty: The moral obligation of non-abandonment at the neural interface. *Journal of Neural Engineering, 6*, 050201 (4pp.).

Fleck, U. (1927a). Über die psychischen Veränderungen der erwachsenen Metencephalitiker mit Betrachtungen über die psychischen Folgezustände der Encephalitis epidemica überhaupt. *Archiv für Psychiatrie und Nervenkrankheiten, 80*, 297–311.

Fleck, U. (1927b). Über die psychischen Folgezustände nach Encephalitis epidemica bei Jugendlichen. *Archiv für Psychiatrie und Nervenkrankheiten, 79*, 723–785.

Foley, P. B. (2003). *Beans, roots and leaves. A history of the chemical therapy of parkinsonism.* Marburg: Tectum.

Foley, P. (2009). Dopamine in perspective. In L. Squire (Ed.), *New encyclopedia of neuroscience* (pp. 563–570). Oxford: Academic.

Foley, P. (2012). The encephalitis lethargica patient as a window on the soul. In L. S. Jacyna & S. T. Casper (Eds.), *The neurological patient in history* (pp. 184–211). Rochester: University of Rochester Press.

Frank, S. A., Wilson, R., Holloway, R. G., Zimmerman, C., Peterson, D. R., Kieburtz, K., & Kim, S. Y. H. (2008). Ethics of sham surgery: Perspective of patients. *Movement Disorders, 23*, 63–68.

Freed, C. R., Zhou, W., & Breeze, R. E. (2011). Dopamine cell transplantation for Parkinson's disease: The importance of controlled clinical trials. *Neurotherapeutics, 8*, 549–561.

Galpern, W. R., Corrigan-Curay, J., Lang, A. E., Kahn, J., Tagle, D., Barker, R. A., Freeman, T. B., Goetz, C. G., Kieburtz, K., Kim, S. Y., Piantadosi, S., Comstock Rick, A., & Federoff, H. J. (2012). Sham neurosurgical procedures in clinical trials for neurodegenerative diseases: Scientific and ethical considerations. *Lancet Neurology, 11*, 643–650.

Goetz, C. G., Bonduelle, M., & Gelfand, T. (1995). *Charcot. Constructing neurology*. Oxford/New York: Oxford University Press.

Guillain, G., & Mollaret, P. (1932). *Les séquelles de l'encéphalite épidémique. Étude clinique et thérapeutique*. Paris: G. Doin & Cie.

Jankovic, J., & Poewe, W. (2012). Therapies in Parkinson's disease. *Current Opinion in Neurology, 25*, 433–447.

Kimmelman, J., London, A. J., Ravina, B., Ramsay, T., Bernstein, M., Fine, A., Stahnisch, F. W., & Emborg, M. E. (2009). Launching invasive, first-in-human trials against Parkinson's disease: Ethical considerations. *Movement Disorders, 24*, 1893–1901.

Lindvall, O., & Björklund, A. (2011). Cell therapeutics in Parkinson's disease. *Neurotherapeutics, 8*, 539–548.

Linke, D. B. (1993). *Hirnverpflanzung. Die erste Unsterblichkeit auf Erden.* Reinbek bei Hamburg: Rowohlt.

Lotmar, F. (1926). *Die Stammganglien und die extrapyramidal-motorischen Syndrome*. Berlin: Julius Springer.

Makowski, J. H. (1983). *Contribution à l'étude des malades mentaux dangereux: itinéraire médico légal et psychiatrique de 1926 à 1982 d'un malade présentant des séquelles d'encé phalite épidémique, ou maladie de Von Economo-Cruchet*. Thesis, Paris.

Master, Z., McLeod, M., & Mendez, I. (2007). Benefits, risks and ethical considerations in translation of stem cell research to clinical applications in Parkinson's disease. *Journal of Medical Ethics, 33*, 169–173.

Mauron, A., & Hurst, S. (2010). Experimenting innovative cell therapies for Parkinson's disease: A view from ethics. In H. Fangerau, H. Fangerau, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 107–122). Bielefeld: Transcript.

Moreira, T., & Palladino, P. (2005). Between truth and hope: On Parkinson's disease, neurotransplantation and the production of the 'self'. *History of the Human Sciences, 18*, 55–82.

Neustadt, R. (1932). *Die chronische Encephalitis epidemica in ihrer gutachtlichen und sozialen Bedeutung*. Leipzig: Johann Ambrosius Barth.

Northoff, G. (1996). Do brain tissue transplants alter personal identity? Inadequacies of some "standard" arguments. *Journal of Medical Ethics, 22*, 174–180.

Northoff, G. (2004). The influence of brain implants on personal identity and personality – a combined theoretical and empirical investigation in 'neuroethics'. In T. Schramme & J. Thome (Eds.), *Philosophy and psychiatry* (pp. 326–344). Berlin: Walter de Gruyter.

O'Sullivan, S. S., Evans, A. H., & Lees, A. J. (2009). Dopamine dysregulation syndrome. An overview of its epidemiology, mechanisms and management. *CNS Drugs, 23*, 157–170.

Politis, M., & Lindvall, O. (2012). Clinical application of stem cell therapy in Parkinson's disease. *BMC Medicine, 10*, 1 (7pp).

Rabins, P., Appleby, B. S., Brandt, J., DeLong, M. R., Dunn, L. B., Gabriëls, L., Greenberg, B. D., Haber, S. N., Holtzheimer, P. E., Mari, Z., Mayberg, H. S., McCann, E., Mink, S. P., Rasmussen, S., Schlaepfer, T. E., Vawter, D. E., Vitek, J. L., Walkup, J., & Mathews, D. J. H. (2009). Scientific and ethical issues related to deep brain stimulation for disorders of mood, behavior, and thought. *Archives of General Psychiatry, 66*, 931–937.

Raja, M., & Bentivoglio, A. R. (2012). Impulsive and compulsive behaviors during dopamine replacement treatment in Parkinson's disease and other disorders. *Current Drug Safety, 7*, 63–75.

Reichardt, M. (1928). Hirnstamm und Psychiatrie. *Monatsschrift für Psychiatrie und Neurologie, 68*, 470–506.

Rosin, H. (1930). Über die Behandlung des Parkinsonismus mit Striaphorin. *Deutsche Medizinische Wochenschrift, 56*, 1046–1047.

Runge, W. (1928). Psychosen bei Gehirnerkrankungen. In: *Die exogenen Reaktionsformen und die organischen Psychosen*. (Handbuch der Geisteskrankheiten, Vol. 7. Spezieller Teil III, pp. 526–700). Berlin: Julius Springer.

Schilder, P. (1922). Einige Bemerkungen zu der Problemsphäre: Cortex, Stammganglien — Psyche, Neurose. *Zeitschrift für die gesamte Neurologie und Psychiatrie, 74*, 454–481.

Steck, H. (1926). Les syndromes extrapyramidaux dans les maladies mentales. *Archives Suisses de neurologie et de psychiatrie, 19*, 195–233.

Steck, H. (1927). Les syndromes extrapyramidaux dans les maladies mentales (suite et fin). *Archives Suisses de neurologie et de psychiatrie, 20*, 92–136.

Steck, H. (1931). Les syndromes mentaux postencéphalitiques. *Archives Suisses de neurologie et de psychiatrie, 27*, 137–173.

Stern, F. (1928). *Die epidemische Encephalitis* (2nd ed.). Berlin: Julius Springer.

Stertz, G. (1921). *Der extrapyramidale Symptomenkomplex (das dystonische Syndrom) und seine Bedeutung in der Neurologie*. Berlin: S. Karger.

Strümpell, A. (1915). Zur Kenntnis der sog. Pseudosklerose, der Wilsonschen Krankheit und verwandter Krankheitszustände (der amyostatische Symptomenkomplex). *Deutsche Zeitschrift für Nervenheilkunde, 54*, 207–254.

Swift, T., & Huxtable, R. (2013). The ethics of sham surgery in Parkinson's disease: Back to the future? *Bioethics, 27*, 175–185.

Thiele, R. (1926). *Zur Kenntnis der psychischen Residuärzustände nach Encephalitis epidemica bei Kindern und Jugendlichen, insbesondere der Weiterentwicklung dieser Fälle*. Berlin: S. Karger.
Trillat, É. (1986). *Histoire de l'hystérie*. Paris: Seghers.
Turner, T. H. (1992). *A diagnostic analysis of the Casebooks of Ticehurst House Asylum, 1845–1890* (Psychological medicine. Monograph supplement, Vol. 21). Cambridge: Cambridge University Press.
Vogt, C., & Vogt, O. (1919). Zur Kenntnis der pathologischen Veränderungen des Striatum und des Pallidum und zur Pathophysiologie der dabei auftretenden Krankheitserscheinungen. (*Sitzungsberichte der Heidelberger Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Klasse*. Abteilung B. Biologische Wissenschaften, no. 14) (56pp).
Volkow, N. D., Wang, G.-J., Fowler, J. S., & Tomasi, D. (2012). Addiction circuitry in the human brain. *Annual Review of Pharmacology and Toxicology, 52*, 321–336.
Voon, V., Mehta, A. R., & Hallett, M. (2011). Impulse control disorders in Parkinson's disease: Recent advances. *Current Opinion in Neurology, 24*, 324–330.
Weintraub, D., & Burn, D. J. (2011). Parkinson's disease: The quintessential neuropsychiatric disorder. *Movement Disorders, 26*, 1022–1031.
Weintraub, D., & Nirenberg, M. J. (2013). Impulse control and related disorders in Parkinson's disease. *Neurodegenerative Diseases, 11*, 63–71.
Zingerle, H. (1936). Über subcorticale Anfälle. *Deutsche Zeitschrift für Nervenheilkunde, 140*, 113–168.

History of Psychopharmacology: From Functional Restitution to Functional Enhancement

30

Jean-Gaël Barbara

Contents

J.-G. Barbara
Université Pierre et Marie Curie, CNRS UMR 7102, Paris, France

Université Paris Diderot, CNRS UMR 7219, Paris, France
e-mail: jean-gael.barbara@snv.jussieu.fr

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_26,

Abstract

This paper presents a very short history of psychopharmacology and attempts to make clear distinctions between (i) the practices of drug administration to insane subjects and the first observations of interactions between diseases and insanity during antiquity, (ii) the concept of psychopharmacology which appears at the end of the eighteenth century with the reform of ancient pharmacology focusing on general knowledge and the closer relations established between chemical substances and mental diseases within the first psychopharmacological rationalisms, and (iii) the rise of psychopharmacology as a discipline after the successes of a few advocates of pharmacological psychiatry in the early 1950s, with the successful uses of chlorpromazine, haloperidol, or reserpine. At each time period, ethical questions arose as to the dangers of these drugs, the addiction risks already being studied in the nineteenth century, the mental diseases which these drugs can induce being studied by means of animal experimentation, and the nonmedical uses of some drugs for enhancement behaviors and better life quality especially in children and students. The epistemology of psychopharmacology as a project arising in the early nineteenth century has developed continuously more complex methodologies for clinical trials. It is a field of intense progress and current inquiry which in return raises new ethical problems to be solved in future decades in order to guarantee progress in biological psychiatry to all social categories.

Introduction

The history of psychopharmacology is a complex historiographical inquiry, since the historical objects in question are often particularly ill-defined and ambiguous. This historical investigation concerns various time periods, from antiquity to the present time, and many scientific disciplines such as medicine, physiology, pharmacology, behavioral pharmacology, neurochemistry, and psychopharmacology. The particular histories published often make the error of comparing projects of distant time periods with disciplines developed later on.

A clear distinction should be made between the ancient pharmacologies of the mind from antiquity to the eighteenth century, the progress of pharmacological chemical synthesis, the evolution in therapeutic uses of drugs, and the various attempts of different periods in clinical trials on insane patients following methodologies progressively worked out with scientific criteria in the nineteenth century. Since each time period prepares the following, the rigorous historical investigation of psychopharmacology of the 1950s must avoid the opposite shortcoming of describing the sudden rise of a new science as being apparently in opposition to anterior knowledge.

Such a study is a sort of challenge as it requires to avoid these numerous pitfalls and forces the historian to define his objects better and distinguish the practices of psychopharmacology (the history of the uses of drugs, antiquity – eighteenth century), the concept of psychopharmacology (the behavioral and physiological study of ancient psychotropic drugs and newly synthesized ones in the nineteenth century), and the discipline of psychopharmacology (since the 1950s).

At first psychopharmacology appears to be characterized by an evident empiricism which raises many questions concerning the ethical issues of clinical trials and the shift from healing therapies to recent enhancement therapies (▶ Chap. 55, "Ethics in Psychiatry"; ▶ Chap. 73, "Research in Neuroenhancement"). However, the epistemological and ethical issues at stake at different time periods are very different, while many histories written by scientists or historians often have the drawback of connecting these issues with technical histories (uses and synthesis of drugs) and the history of ideas in a similar manner.

The modern historiographical approach to psychopharmacology should aim at working out these issues and better defining the limits of historical continuities and the real changes in each period, concerning not only science itself but especially epistemology and ethics. The present study only represents a short introduction to such a project hopefully to be expanded in the future.

From Antiquity to the Eighteenth Century

Scipion Pinel, the son of French nineteenth-century alienist Philippe Pinel, condemned the psychopharmacological practices of past centuries and their false empirical character which justified the use of drugs not by experience but by bringing into play outdated medical theories: " [. . .] One must regret the blind belief in disruptive medicines [based on past theories] [. . .] [These theories] are at times the coction of humours, the revulsion and repulsion of peccant matters in the brain, which these impatient doctors fight with powders, extracts, juleps, electuaries, potions, to defeat insanity; At other times, they prescribe bloodletting excessively, with no distinction of the causes and epoch of the disease, showing a medical doctrine full of prejudice, pedantry and ignorance, at various periods, often supported by the disastrous association of the name of a famous author." (Pinel 1837, p. 132; author's translation).

If therapeutic empiricism seems evident in the nineteenth and twentieth centuries, the psychopharmacological empiricism before the eighteenth century was mixed with the sophisticated doctrines of the giant medical systems.

Scientific psychopharmacological empiricism, such as the empiricism claimed by Claude Bernard, being used to combat these medical rationalisms and systems which evaded experience and experimentation, cannot easily be found in the periods previous to the nineteenth century, although local drug experimentation on the insane is not absent from antique and Arabic medicine.

What is often missing before the nineteenth century is the pharmacological empiricism based on novel psychopharmacological rationalisms which we define as the convergence between the development of new theoretical ideas on mental diseases, new rational explanations of the actions of drugs, and new practices of drug therapies. In the eighteenth century these convergences were often based on relations between outdated theories (theory of humors), past and new practices with weak evaluation of treatments, and poor attempts to consider new explanations of drug actions.

This is, in my opinion, the way one should consider psychopharmacology from antiquity to the eighteenth century, as it presents great geographical and chronological heterogeneity in this time period. Nevertheless, I will only consider here a few continuities in the psychopharmacological practices.

Greek medicine failed to build strong correspondences between a unified etiological conception of mental illness and a set of treatments tested with empirical therapeutic trials. However, the Hippocratic corpus described the organic etiology of some mental diseases, such as the "sacred disease" (epilepsy), and put forward the theory of humors to explain it, thus leading the way to possible drug treatments able to reduce the excess of phlegm in the brain as observed in the "epileptic goat."

In this area of medicine, Greek doctors gathered interesting clinical observations, noticed and studied further in the following centuries, in order to imagine new ways of treating mental diseases based on some interactions between insanity and pathologies such as fevers which sometimes improved mental health. These observations led to pharmacological treatments up to the twentieth century.

Therefore, if the concept of psychopharmacology was absent before the nineteenth century, a psychopharmacological thinking was on the way locally, in a nonsystematic but positive manner. This way of reasoning is relevant to our inquiry since it is in continuity with the treatment of the insane in the following centuries and because it remained heuristic until the rise of the concept of psychopharmacology in the nineteenth century.

The Concept of Psychopharmacology (Nineteenth Century)

The Rational Refusal of Drugs

At the end of the eighteenth century, Philippe Pinel (France), William Tuke (Great Britain), Vincenzo Chiarugi (Italy), and Benjamin Rush (United States) attempted to improve the conditions of confinement of the insane, leading the way to moral treatment. This turn came with a mainly moral conception of mental illness, although these doctors did not ignore, whenever possible, lesional causes and interactions with organic pathologies. Therefore, Pinel, for example, did not refrain from prescribing medications to improve the health of his patients. He himself used a short pharmacopoeia with favorite remedies such as wild blackberry and a mixture of camphor (Guislain 1826).

However, these different types of drugs of the early nineteenth century were rather inefficient, as was the massive use of herb tea in asylums until the first half of the twentieth century.

Scipion Pinel himself did not deny the value of medications, but the "monstrous polypharmacy" (his term) of the insane asylums was largely disastrous in his opinion, as was the immoderate use of bloodletting and the purgatives of ancient medicine.

Novel and Often Naive Rationalisms

Although insane asylums kept some specific traditions of old medications in very contrasted manners, with large differences of practices and opinions concerning the efficiency of drugs (Curchod 1845), a novel rationalism burst out to the scene at the turn of the nineteenth century, within a context where the conception of mental illness was divided into a moral etiology and an organic etiology.

In parallel to the rise of experimental psychology as a science as early as the 1820s (François Magendie, Pierre Flourens, or Luigi Rolando), some alienists with anatomopathological background, and those practicing autopsies, were convinced of the organic etiology of insanity, especially after Antoine Laurent Bayle (1799–1858) managed to demonstrate the organic and syphilitic origin of the general paralysis associated with insanity. These doctors adopted a complex epistemology of the anatomopathology of mental diseases, in line with Franz Joseph Gall, which demanded that primitive organic lesions should be searched for systematically in the autopsies of insane subjects (Barbara 2011a). They refused the concept of "pure functional lesion" which, however, surfaced at the end of the nineteenth century in Paris during the famous polemics on hysteria in the circles of Jean-Martin Charcot (Barbara 2011b).

Some of these alienists, such as Jacques-Joseph Moreau de Tours, developed interesting parallels between the organic action of chemical substances, such as cannabis, and insanity. When back from a journey in the East, Moreau de Tours made self-experimentation with cannabis and described the dreams associated with the drug which he equated with symptoms of insanity. Such a parallel was built over the central idea of the unicity of the etiology of nervous diseases and insanity, which Moreau de Tours suggested after his observations of the nervous and psychic effects of cannabis (Ledermann 1988). For Moreau, cannabis became a tool for experimental pathology, in the line of experimental physiology or, in Moreau's terms, of "mental pathogeny." Chemical substances became, as they were during or the psychopharmacological revolution of the 1950s, a means to build and discover new etiologies of mental illness. More specifically, for Moreau, the drug (cannabis) acted by way of a "substitutive action" on the cause of insanity (the "generator fact" of Moreau). The action of the drug restored the perturbation induced by the mental defect. Such a toxicological etiology of insanity can be seen as naive, but it represents a novel psychopharmacological way of thinking and a heuristic path to discovery opened by experimental physiology still widely used today. In the nineteenth century, a central theme of this strategy was the study of the mechanisms of poisons, such as the "nervous poison" curare (Barbara 2009).

New Drugs and New Clinical Trials

As early as the very beginning of the nineteenth century, new active plant alkaloids were chemically isolated and used in medicine: morphine, strychnine, caffeine, quinine, veratrine, atropine, or cocaine. Then the era of chemical

syntheses started with the famous leaderships of French, and then German chemists; the preparations of new bromides; and the chemical synthesis of chloral, chloroform, barbiturates, and paraldehyde. Novel animal experimentations tested the physiological actions of new and old drugs, such as Indian poisons used for hunting. In a second perspective, therapeutic trials were progressively widened to include insane subjects.

Pharmacology, experimental physiology, and novel toxicology were the scientific disciplines concerned. Before 1806, French physiologist, François Magendie, studied the physiological actions of upas and, some 10 years later, he experimented on its active principle, strychnine, just after its chemical isolation. As early as 1826, Pierre-Alexandre Charvet presented his dissertation entitled *Proposal on the mode of action of Opium in man and animals* at the French medical faculty. Deguise, Dupuy, and Leuret published their research on morphine in 1824 based on physiological and anatomopathological inquiry on animals (Chap. 134, "Animal Research and Ethics"). In the 1840s, Flourens also worked with a similar perspective on ether and chloroform. Experimental physiologists progressively defined these drugs as "nervous poisons," while they interpreted their actions with the scientific knowledge of nerve physiology (Barbara 2010).

All these studies did not proceed from a blind empiricism, although chance was a key factor in gathering Indian curares and upas. A profound knowledge of the power of these ancient poisons was scientifically used and tested experimentally and methodically. At the same time, rigorous clinical trials were developed from the end of the eighteenth century with old drugs, such as opium and morphine, and with a new concern in rationalizing pharmacy which tended to escape general knowledge and to develop regionally. Rational pharmacological trials can be found, however, earlier in the writings of some medieval and Arabic physicians. But the therapeutic indications of drugs varied greatly and agreements can hardly be found: before ether was used for anesthesia, it was used for gout, headache, or gastric spasms.

Therapeutic trials on the insane were made in some asylums, but few concerning drugs, and more generally involving moral treatments and physical therapies.

Some physicians experimented on the new sedative properties of the drugs studied by physiologists, directly on insane subjects, such as Thomas Smith Clouston (1887) who defined some basic scientific principles, some of them using statistics (Clouston 1863). Like Moreau de Tours, other physicians developed physiological and pharmacological ideas on insanity which justified their experimentations, such as the clinical trials of ether inhalation on epileptic patients in the 1840s.

The relations outlined by Greek doctors in antiquity between mental illness and diseases led to organicist conceptions of mental diseases favoring medications. German physician Wilhelm Griesinger widened the conception of a neuropathic origin of mental diseases which led to the idea that, if a central element of the nervous system is altered, medications can act on the nervous peripheral origins and fight the pathological processes invading the brain.

These approaches aiming at a more scientific rationalism compared to the previous decades, especially concerning the use of group subjects, control subjects, averages, and statistics, led to very little success, but to a better knowledge of sedation with available drugs until the 1940s.

However, during the 1860–1880s, physicians refined their practices of drug administration and developed different treatments for each nosographical category of mental diseases (Kraepelin 1899). Physicians generally followed the first recommendations on the cautious uses of drugs by François Magendie and viewed problems of toxicity and addiction more objectively (▶ Chap. 64, "What Is Addiction Neuroethics?"; ▶ Chap. 67, "Ethics Issues in the Treatment of Addiction").

Experiments and Newly Synthesized Drugs at the Turn of the Twentieth Century

The Introduction of Experimental Psychology in Psychiatry

At the turn of the twentieth century, a new scientific discipline, namely, experimental psychology, took over the objective inquiry of the effects of drugs on the mind by members of psychological faculties, in particular those interested in physiology and psychiatry, such as Kraepelin, Binet, and Féré.

In the 1880s, alienists, such as Clouston, claimed that medical psychology should help psychiatry in its new physiological perspective, thus evading the single previous philosophical framework that made psychology a branch of philosophy (Clouston 1887). This era marked by experimental physiology enabled experimental psychology and psychiatry to claim legitimacy for their medical investigations of the mind, evading the purely spiritualist perspective and taking the lessons of experimental psychologists such as Kraepelin, who trained with Wilhelm Wundt on the measure and interpretation of reaction times.

One of the new experimental approaches used Mosso's ergograph to study muscular fatigue and psychism, when a muscle is exhausted by intense work and nevertheless recruited again by a surge of lasting psychical force. This type of work was taken over by experimental psychologists and adopted in the study of the effects of drugs on psychism, in particular in learning and memory (Kraepelin, Benedict; see Lashley 1917). This was the way to the concept of behavioral pharmacology, a discipline adopting reflex tests, capacity tests of psychic faculties, control subjects, and simple statistical analyses.

The New Project of Psychobiology

Psychologists, alienists, and physiologists, individually or collectively and interdisciplinarily, not only claimed the experimental study of the mind, but they also advocated and tried novel forms of cooperation between disciplines in a new interdisciplinary perspective.

The term “psychobiology” never became the label of a clearly defined approach to psychology, but nevertheless, in the early twentieth century, it designated a diversity of opinions on what the relationship between experimental psychology and biology should be (Dunlap 1914).

According to Dewsbury, such opinions all shared a common point of view in their wish to bring psychology closer to biology in a holistic vision of the organism, in a manner analogous to the wish of biologists such as neurophysiologists Ralph Gerard (USA) or Alfred Fessard (France) to always keep in mind a “large picture” of the scientific problems, while biology turned to the systematic study of the elementary mechanisms of nervous phenomena (Edgar Adrian) and the minute molecular structures (Jean Nageotte, Francis O. Schmitt) (Dewsbury 1991).

Therefore, psychobiologists expressed the parallel wishes to preserve psychology from biological reductionism adopted uncritically, while taking advantage of the advances of biomedical disciplines. However, these psychobiological projects never became united, and were often aborted, such as the journal *Psychobiology* of K. Dunlap, with only two issues published in 1917 and 1920. Some of these projects, however, showed a great anticipation of the spirit of the neurosciences, as for example in France, where Henri Wallon created a *laboratoire de psychobiologie de l'enfant* (laboratory of child psychobiology) in a primary school in the Parisian suburbs in 1925 (Galifret 1979).

One can clearly see, in the perspective of Henri Wallon or that of K. Dunlap, the wish to take advantage of the advances in the study of the “histological details” in the field of psychology and in the desire to explicate their psychological and functional significance in the classical perspective of “physiological psychology.” In a broader perspective, those researchers claiming a psychobiological project did not put up any opposition between the study of forms and that of functions. These theoretical opinions favored studies on the effect of drugs on the psychic functions of normal subjects and diseased patients.

The new orientations of psychology in the 1920s and 1930s offered a new ground for interdisciplinary projects with biologists, physiologists, and physicians, some of them massively funded by the Rockefeller Foundation.

Psychopharmacology as a Project

In the 1920s, one can observe a burst of new studies on the effects of drugs on psychic faculties in animals and man. In this psychobiological context, David Israel Macht (1882–1961), pharmacologist of Russian origin at Johns Hopkins University, started experimental research on behavioral pharmacology on mice, developing a motor skill using a rope or circular mazes (Macht and Mora 1920). In one of his articles, he designated his emergent research program, in the line of the previous studies of experimental psychology of the 1890s, under the label “psychopharmacology,” although the term – or similar ones – had previously been used by others (Macht 1920). Macht defined his investigations very simply

as the study of the effects of drugs on psychological functions, with nothing really new from the previous psychopharmacological concept.

The interdisciplinary character of the study of Macht, and of others at the same time, clearly appeared since those projects involved not only physicians but also physiologists and pharmacologists. This is in striking opposition to the *Pharmakopsychologie* of Kraepelin which largely remained programmatic (Kraepelin 1892).

The Empiricism of Shock Therapies and Pharmacological Cures

In parallel to the development of a scientific project in psychiatry, a new path of therapeutic empiricism emerged at the turn of the twentieth century, aiming at understanding the effects of drugs on the mind.

Classical histories of biological psychiatry often focus on the first very successful psychiatric treatment: the malaria therapy of von Jauregg. However, the success of the treatment of neurosyphilis by inoculation of a malaria clone of low virulence in order to provoke fever represented the end product of a very long series of trials to induce a great variety of diseases, some of them dangerous, and to alleviate insanity.

One should reconsider those efficient clinical trials in the larger framework of the medical empiricism used in healing insanity, where physical and chemical treatments of all sorts were considered, such as the prolonged narcosis with bromide prescribed in order to reorganize functional nervous activities. This empiricism was characterized by the poor ethical concerns of trials and the lack of means to evaluate the benefit of treatments (▶ Chap. 58, “Relationship of Benefits to Risks in Psychiatric Research Interventions”). But scientific justifications are nevertheless elaborated with reference to a biomedical rationalism still based on nervous physiology and the emerging neuron doctrine.

Naive Rationalisms of Shock Therapies and Pharmacological Cures

The histories of shock therapies and prolonged pharmaceutical treatments used in psychiatry refer to other rationalisms responsible for the choice of inducing agents but also to the scientific explanation of their positive effects which justify their use – with little proof however.

The deep sleep therapy of Macleod and that of Epifanio using barbiturates and of Klaesi are all based on the belief that putting a stop to psychic activities in the nerve centers enables a better start upon wakening and the associated functional reorganization of neuronal connections, which explained the improvement of the observed psychic state of the patients, in a similar manner to the sleep theory proposed by Matthias Duval in the nineteenth century.

Hungarian anatomopathologist Meduna, trained in psychiatry, developed a psychobiological rationalism based on his histological investigations.

His observations of glial cells on epileptic and schizophrenic patients led him to think that if epilepsy induces a glial reaction, those cells tend to disappear in patients with schizophrenia. Meduna inferred that, in schizophrenic patients, glial cells could be reactivated with the artificial induction of epileptic convulsions. For this purpose, he used the intravenous administration of a convulsant, camphor, which led him to the rationalism of the Cardiazol treatment and convulsive therapy.

Sakel also used this rationalism at the turn of the twentieth century, along with the doctrine of Ramón y Cajal, in order to explain the benefit of the insulin treatment where the induction of a coma was believed to stop the degenerating processes accounting for altered cells and to enable the regeneration of their connections.

These types of reasoning were not much more elaborated than those of the malaria therapy, which leads epistemologists to infer that those researches were profoundly empirical. However, their empirical character rather results from ad hoc explanations, based on theoretical frameworks with mediate relations to the facts observed and which remain largely hypothetical.

The Beginning of Psychopharmacological Therapies

Limits of the Classical Pharmacology of the Mind

The pharmacology of mental diseases still remained very unsatisfactory until the 1950s. Some progress was made during the 1940s, but pharmacological innovation often continued to be homemade in small family industries, in spite of the development of large pharmaceutical groups at the international level.

The pharmacology of the mind was a poorly developed branch of psychiatry which nevertheless tried to make legitimate and somewhat efficient uses of narcosis, psychosurgery, and shock therapies on patients, with the aim of sedating them to secure psychotherapy and occupational therapy.

This failure to create new pharmacologically efficient treatments explains the impossibility of psychiatrists to find any consensus in the explanation of the effects of drugs and in psychogenic theories of mental diseases. This situation was more or less the one that prevailed over the previous centuries.

Locally, research groups managed to improve the efficiency or reduce the toxicity of some drugs, as was the case for specific classes of antiepileptic drugs, such as hydantoins. LSD or low doses of curare were tested on human subjects during the Second World War and on patients with mental diseases, while shock therapies still remained in wider use in asylums.

However, in the 1940s, some psychiatric departments embarked in therapeutic pharmacological empiricism (Jean Delay in France, Joel Elkes in Great Britain), after the rise of pharmaceutical industries opened an era of great enthusiasm in medicine, with dominant economical positions acquired during the Second World War, and following the discovery and massive use of new antibiotics, steroids, and antituberculosis drugs.

Therefore, the 1930s and 1940s should not be considered as a nonprofitable period for pharmacology and psychiatry, although the great success of neuroleptics in the 1950s overshadowed the new prescriptions of barbiturates, which replaced bromides, and the introduction of antiepileptics and anticonvulsants, such as phenytoin and trimethadione.

Psychiatry and Pharmaceutical Innovation

In the 1940s, a few neurologists and psychiatrists started experimentations on animals and new clinical trials on their patients, with little ethical commitments, trying new drugs and sometimes modifying the chemical structure of molecules themselves.

New effects of drugs unknown until then, but suggested by scarce and incomplete observation, were tested on specific categories of patients, as was the case for hydantoins, LSD, and disulfiram.

Past therapeutic empiricism met with a novel enthusiasm. Such empiricism had been forgotten when the therapeutic properties observed in the newly synthesized drugs were often different by far from those expected.

The path to discovery of the antihistaminic drugs included the phenothiazines previously recognized as insecticides, anthelminthic drugs, and then antimalaria drugs. The French military surgeon, Henri Laborit, experimented on the action of the antihistaminic properties of drugs in the potentiation of anesthesia to reduce anesthetic agent doses and prevent postsurgical shock and anesthetic shock.

Novel Psychopharmacological Rationalisms and the Neuroleptic Concept

New phenothiazines, the drugs which were to become later Antergan®, Neoantergan®, Phenergan®, and Multergan®, were synthesized in France in the 1940s and showed interesting antihistaminic properties. Phenergan® was tested at Sainte-Anne hospital in Paris by Dr. Paul Guiraud, as a sedative and hypnotic drug.

The fabric of the neuroleptic concept and the making of chlorpromazine as a drug against psychoses was a tortuous path pursued when Laborit, joined by French pharmacologist Pierre Huguenard, introduced a lytic cocktail with Phenergan®-Dipacol® which was shown to induce “artificial hibernation” and used to protect wounded soldiers from traumatic shock during the French Indochina War.

In these studies, Laborit tried a promethazine from the French pharmaceutical industry, Rhône-Poulenc, namely, chlorpromazine. Laborit rapidly noticed the antipsychotic action of the molecule which nevertheless remained unnoticed by psychiatrists from Val-de-Grâce military hospital where Laborit worked. Chance was partially involved in the discovery by Pierre Deniker and Jean Delay at Sainte-Anne hospital of the antipsychotic properties of chlorpromazine used alone,

at a time when drugs were generally used in combination or along with other physical treatments, such as sedation by lowering the body temperature with ice. The discovery occurred when the sedation with ice was abruptly interrupted, without notice to the doctors, because of an ice shortage in the hospital, but with no consequence on the benefit of the chlorpromazine treatment.

The Clinics of Chlorpromazine

The great, but slow, worldwide success of the discovery of the neuroleptic effect of chlorpromazine triggered enthusiasm among some pharmacologists and psychiatrists. The rationalism of antihistaminic drugs with expected antishock properties, given that they prevent anaphylactic shock, could lead other researchers to use a similar simple reasoning to find new psychiatric medications.

Delay visited laboratories abroad and was actively involved in the diffusion of his discovery of the effects of chlorpromazine. The molecule was progressively adopted in many countries and radically changed the life of psychiatric departments when the sedation of patients allowed social therapy.

Chlorpromazine slowly became a neuroleptic medication, but for numerous years after its discovery, it remained advertised, commercialized, and prescribed only as a sedative drug. The sedative chlorpromazine replaced hyoscine or amytal and led, however, to the withdrawal of polytherapies, since it was efficient when used alone. Chlorpromazine as a neuroleptic drug required many international meetings and international scientific prizes to advertise the new medication as such.

Therefore, one should always remember that a medication possesses an essence radically different from that of its constituting molecule. Chlorpromazine made possible the manufacture of Largactil® as a neuroleptic medication after 10 years of use worldwide.

Psychopharmacology as a scientific and medical discipline was born from this slow revolution, when a new drug was discovered to be efficient alone, and this ultimately paved the way to new explanations of psychosis. The revolution did not involve any radical change in pharmacological rationalism, but in the status of psychiatry in medicine and society with the new and markedly recognized efficacy of psychiatrists.

New Frameworks of Psychopharmacology and Theories

The discovery of haloperidol by Paul Janssen followed a rationalism different from that of chlorpromazine, but as simple and efficient, which was described by philosopher and historian Jean-Noël Missa after an interview with Janssen (Missa 2006). A friend of his, practicing high-level biking, informed him of the psychogenic effects of the amphetamines used for doping. Janssen felt he could discover anti-amphetamine drugs with possible antipsychotic effects, and this is how he managed to synthesize and test haloperidol. As for chlorpromazine, the new

molecule was first adopted as a sedative drug, and Janssen had to fight and convince psychiatrists to search specifically for the antipsychotic properties of the Largactil® commercialized form.

In this context, many psychiatrists tried their luck at testing old drugs or new molecules, such as reserpine, LSD, monoamine oxidase inhibitors, or benzodiazepines.

Psychiatrists, physiologists, and pharmacologists joined into common projects. In the USA, Seymour Kety became director of a new and important interdisciplinary psychiatric center at NIH. Those scientific interactions enabled common interpretations of the new efficient treatments and the rise of new biological theories of mental diseases based on the depletion of neurotransmitters (catecholaminergic theory of depression by Axelrod, serotoninergic theory of depression by George Ashcroft and Donald Eccleston) or the hypersensitivity to others (dopaminergic theory of psychosis by Carlson).

These new theories appeared a few years after the great controversy surrounding the chemical theory of neurotransmission and represented an extension of the latter in the medical field.

Psychopharmacology as a Scientific Discipline

Psychopharmacology as a discipline was born from a small group of psychiatrists, pharmacologists, neurochemists, one neurologist, and one surgeon who, in the 1940s, chose to follow the path of pharmacological empiricism to treat mental illness.

Only after the early successes of the 1940s, and mainly the 1950s, did this little community gather in international congresses and decided to found a new society, the *Collegium Internationale Psychopharmacologicum*.

From the very early 1950s onwards, they organized interdisciplinary meetings on psychiatric drugs and their mode of action (Jean Delay, 1950, Paris; US NAS, Seymour Kety, 1954; NIH, 1956, Ralph Gerard; Macy foundation, 1954–1959, Harold Abramson).

Following the introduction of radiolabelled neurotransmitters and the rise of neurochemistry, biological psychiatry, psychopharmacology, neurochemistry, and neuropharmacology became much more than individual disciplines. They became real scientific networks overlapping and paving the way to new forms of interdisciplinarity in the underground neuroscience project on its way in the late 1950s.

Ethics of Psychopharmacology

Ethics of Clinical Trials and Drug Evaluation

Chlorpromazine and haloperidol were tested on the patients of psychiatric departments with few ethical procedures, at a time when shock therapies were still in use, in spite of the Nuremberg code which was not strictly applied and in a scientific way not very far from standard physiological experimentation.

In September 1956, in Washington D.C., a meeting focused on psychopharmacology and the problems in evaluation, under the presidency of Ralph Gerard. The main issues concerned scientific and ethical questions dealing with patient categories, the evaluation of treatment benefit, adequate statistical analyses, conflicts of interest, the risk of side effects, problems of the autonomy of subjects, and the equity of access to treatment among patients.

While psychopharmacology was progressing, local ethical committees were constituted in institutions in order to guarantee an evaluation independent of the actions of pharmaceutical companies. Historians and sociologists of science have shown how mental diseases and their treatments could be seen as social constructs dependent on the economy, the social characteristics of nations, the interests of pharmaceutical industries, and the degree of the infiltration of psychiatry by psychoanalysis.

Questions on the autonomy, consent, and the selection of patients are still complex today in the fields of psychopharmacology (Informed Consent in the Mentally Ill). A controversy still occurring today concerns the use of electroconvulsive therapy. Although the techniques have been used for decades and advocated as a safe method particularly suited to treat psychotic depressions, some patients report traumatizing memory loss. For each therapy specific issues should guarantee an equitable negotiation of the consent form between patients and doctors.

In order to test the negative side effects of treatments and their unsuspected psychological consequences, some psychiatrists have practiced autoexperimentation with electroshock. Those experiments have shown to the medical community deleterious psychological consequences, more pronounced than those which could be evaluated by simply recording the improvement in the psychiatric state of the patients. These practices helped to justify the inclusion of healthy subjects in clinical trials, and since 1976 the American College of Neuropsychopharmacology enounced some restrictions on the conditions concerning the inclusion of such subjects.

Enhancement

In the previous years, sociologists and philosophers have studied the social consumption of psychiatric drugs for enhancement of cognitive capacities, in particular among children and teenagers during schooling (▶ Chap. 75, “Neuroenhancement”; ▶ Chap. 80, “Reflections on Neuroenhancement”; ▶ Chap. 73, “Research in Neuroenhancement”).

In a more general manner, nonmedical psychiatric drug consumption is an old theme of nineteenth-century medicine with the denunciation of addiction risks (▶ Chap. 64, “What Is Addiction Neuroethics?”).

People are deceived when some drugs are advertised, by pharmaceutical companies and society, as efficient treatments to a selective health problem. They consider science as a means to improve their quality of life but with no critical analysis, which should be adequately provided by professionals, physicians, drugstores, and drug makers.

France is the leader in antidepressant and anxiolytic drugs. Modafinil® is used in the US against narcolepsy and mild sleep disorders to enhance cognitive capacities.

The consumption of Ritaline® by children seems medically justified in only about 10 % of the subjects (▶ Chap. 106, "Neuroethical Issues in the Diagnosis and Treatment of Children with Mood and Behavioral Disturbances"; ▶ Chap. 73, "Research in Neuroenhancement").

Philosophers and ethicians tend to broaden the debate on these dangerous behaviors and present enhancement as a long-standing cultural reality. What really matters is the risk of selling drugs to consumers to combat, for example, memory loss, in the same manner as aspirin is provided against fever. The concepts of drug and enhancing agent are close and the boundary between the two is sometimes difficult to see for users who feel the right to correct minor impairments with pharmacological treatments without the help of responsible, not hypocritical, physicians.

Future Questions

The ethics of psychopharmacology remains a vast and often difficult domain of inquiry because of the dilemmas occurring on account of the application of antagonist principles of clinical trials which must simultaneously respect the autonomy of the subjects and guarantee a benefit, while giving access to new treatments to all social categories.

One of the major issues at stake deals with the trials of underrepresented social categories for which obtaining consent is complex, as for children (minors), teenagers (respect of their autonomy), pregnant women (risks for the baby), and minorities (possible lack of confidence and understanding). An ethical controversy concerns random trials with placebo controls using randomized withdrawals of patients. This is in opposition to the benefit principle, but it remains a key condition to evaluate this benefit.

For these reasons clinical trial methodologies become more and more complex and tend to be based on large international networks with rigorous follow-up meetings. These networks should remain independent, which is a difficulty given the necessary funding by industries and possible conflicts of interests.

The ethics of psychopharmacology is still an advanced domain of bioethics where much research and negotiation are still necessary. Historians of science are not excluded since a historical perspective can show very recent abuses and instrumentalizations of psychopharmacological knowledge by politicians, military institutions, and industries.

Acknowledgement The author thanks Chantal Barbara, Jean-Noël Missa and Frank Stahnisch for helpful comments and suggestions.

Cross-References

▶ Research in Neuroenhancement
▶ What Is Addiction Neuroethics?

References

Barbara, J.-G. (2009). Claude Bernard et la question du curare : Enjeux épistémologiques. *Journal de la Société de Biologie, 203*, 227–234.

Barbara, J.-G. (2010). *La naissance du neurone*. Paris: Vrin.

Barbara, J.-G. (2011a). Ouvrir le corps des fous et des criminels : Science et enjeux philosophiques d'hier et d'aujourd'hui. In *Crime et Folie*, L. Bossi (Ed.), *Les cahiers de la NRF* (Entretiens de la fondation des treilles). Paris: Gallimard.

Barbara, J.-G. (2011b). Relations médecine – sciences dans l'individualisation des maladies à La Salpêtrière à la fin du XIXe siècle. *Revue d'Histoire des Sciences, 63*, 369–407.

Clouston, T. S. (1863). *Contributions to the study of insanity and its treatment*. Garlands, Carlisle: The Cumberland and Westmorland asylum.

Clouston, T. S. (1887). *Clinical lectures on mental diseases*. London: Churchill.

Curchod, H. (1845). *De l'aliénation mentale et des établissements destinés aux aliénés dans la Grande Bretagne*. Lausanne: Bridel.

Deguise, M., Dupuy, A. C., & Leuret, F. (1824). *Recherches et expériences sur les effets de l'acétate de morphine*. Paris: Crevot.

Dewsbury, D. A. (1991). Psychobiology. *American Psychologist, 46*, 198–205.

Dunlap, K. (1914). *An outline of psychobiology*. Baltimore: The Johns Hopkins Press.

Galifret, Y. (1979). Le biologique dans la psychologie de Wallon. *Enfance, 32*, 355–362.

Guislain, J. (1826). Traité sur l'aliénation mentale et sur les hospices des aliénés. Amsterdam, J. van der Hey et fils, Les héritiers H. Gartman.

Kraepelin, E. (1892). *Über the beeinflussung einfacher psychischer vörgange durch einige arzneimittel* (p. 227). Jena: G. Fischer.

Kraepelin, E. (1899). Psychiatrie: Ein Lehrbuch für Studirende und Ärzte Leipzig: Verlag von Johann Ambrosius Barth.

Lashley, K. S. (1917). The effect of strychnine and caffeine upon the rate of learning. *Psychobiology, 1*, 141–170.

Ledermann, F. (1988). Pharmacie, médicaments et psychiatrie vers 1850: Le cas de Jacques-Joseph Moreau de Tours. *Revue d'Histoire de la Pharmacie, 276*, 67–76.

Macht, D. I. (1920). The effect of some antipyretics on the threshold of hearing. *Journal of Pharmacy and Experimental Therapeutics, 15*, 149–165.

Macht, D. I., & Mora, C. F. (1920). Effect of opium alkaloids on the behaviour of rats in the circular maze. *Journal of Pharmacology and Experimental Therapeutics, 16*, 219–235.

Missa, J.-N. (2006). *Naissance de la psychiatrie biologique*. Paris: PUF.

Pinel, S. (1837). *Traité complet du régime sanitaire des aliénés*. Bruxelles: Société encyclographique des sciences médicales.

Informed Consent and the History of Modern Neurosurgery

31

Delia Gavrus

Contents

Abstract

The concept of consent has a long history, being defined and understood in different ways over time. Specific historical events and debates, both internal and external to the medical profession, have shaped the manner in which experimentation on humans has been conducted and regulated. The adoption of laboratory approaches and an alliance with experimental science in the nineteenth century transformed the medical profession. Despite early attempts to articulate an ethics of experimentation, invasive and dangerous therapeutic procedures, including on the brain, were often tried on patients without their consent. The Nuremberg Code set important criteria, but egregious experimentation continued after the war. It was in the 1960s and 1970s, during the rights movement, and in the wake of public outrage over a number of revelations that a modern notion of informed consent began to take shape and that medical practices became subject to some external safeguards and regulations. The history of neurosurgery mirrors this more general story. After two decades of bold attempts to open the skull for a variety of afflictions, the development of neurosurgery as a medical specialty in the early twentieth century brought with it

D. Gavrus
Department of History, University of Winnipeg, Winnipeg, MB, Canada
e-mail: delia.gavrus@utoronto.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_24,

an initial conservatism both in technique and in the range of conditions for which brain surgery was attempted. Starting in the 1930s, a radical interventionist ethos paved the way for the wide adoption of procedures such as lobotomy. A thorough scrutiny of these procedures came only in the 1970s, when review boards and other safeguards were put into place.

Introduction

As a practice, vocation, and profession, medicine has long been engaged in a spirited dialectic about ethics with society at large. The doctor-patient relationship pivots on an imbalance of power; issues of authority and trust shape the outcome of the therapeutic encounter, and over time doctors, patients, and regulators have responded in historically contingent ways to a complex set of challenges that pits investigation against care, healing against experiment. Focusing on the United States, in part because it was here that neurosurgery first developed as an organized medical specialty, this chapter will offer a brief review of the literature on the history of human experimentation, medical ethics, and informed consent in the nineteenth and twentieth centuries, and it will sketch the history of modern neurosurgery in relationship to these issues. Since neurosurgery involves the handling of the brain – the organ that in modernity is most closely associated with the self – this medical specialty has raised and will likely continue to raise deep questions about the ethics of treatment, especially in those cases in which cognitive and emotional impairment calls voluntary informed consent into question.

Human Experimentation and Informed Consent

Historians have argued that in the second half of the nineteenth century, the adoption of laboratory approaches and an alliance with experimental science transformed the medical profession, affecting not only the practice of medicine but also doctors' professional identity, their relationship with patients, and the public perception of medicine (Cunningham and Williams 1992; Schlich 2007; Shortt 1983; Warner 1986, 1991). Experiments had been conducted occasionally before this time (Trohler 2007), but doctors now started to view experimentation on patients and healthy individuals as highly desirable and necessary for medical progress, and an increasing number of medical journals and institutions began to cater to this enterprise. A concern with the ethics of experimentation was present at this time. The French physiologist Claude Bernard, who had published the influential book *An Introduction to the Study of Experimental Medicine* (1865), insisted that experiments that could harm a person should never be performed, while several decades later the celebrated physician William Osler emphasized not only safety, but also the consent of the patient (Rothman 1996).

However, despite these high-profile views, not only were invasive and dangerous therapeutic procedures tried on patients, but an enthusiasm for the new science

of bacteriology led some late nineteenth-century doctors to infect individuals (including themselves, orphan children, and those deemed to be "feebleminded") with syphilis, gonorrhea, smallpox, and other pathogens. Similar experiments were performed with thyroid extract, cancer grafts, and techniques such as lumbar puncture. The patients sometimes assented to the experiment; in many other cases they needed considerable persuasion, and in other cases they were deliberately led to believe that the procedure was part of the treatment. Physicians like Osler voiced disapproval when a few particularly egregious cases drew public attention, but for the most part, these kinds of studies continued to be performed and published in the medical literature. In the United States, reform-minded citizens raised concerns about such practices, and a lively antivivisectionist movement embraced what it considered to be the twin goals of animal and human protection from immoral experimenters. To the distress of the American Medical Association (AMA), which preferred to address the control of such practices internally, the antivivisectionists lobbied for federal legislation to regulate animal experimentation, an effort that led to Senate Committee hearings in 1900. The same year, a bill to regulate human experimentation was introduced but failed to become law. The medical profession strongly defended their research practices against the antivivisectionists, and, while rejecting external oversight, a few proposals for guidelines for clinical investigators and animal experimentation were put forward at the beginning of the twentieth century. The AMA code of ethics, however, was not revised to require voluntary consent until the 1940s (Lederer 1995). In other countries such guidelines were available earlier, although enforcement was a complicated issue. For instance in Germany, 1931 regulations that governed the introduction of new therapies emphasized informed consent, but such regulations did not affect the atrocious medical experimentation during the war.

In principle nineteenth- and early twentieth-century physicians subscribed to the idea that patients had to be informed before a dangerous procedure was undertaken. Nevertheless, this impulse to be truthful was balanced by a belief that it was a physician's duty not to divulge information that could unduly distress the patient and worsen his or her condition. Such benevolent deception had been enshrined since 1847 in the AMA code of ethics and had been heavily influenced by the work of the English physician Thomas Percival. At the beginning of the century, Percival had published a book that addressed medical conduct and ethics, emphasizing gentlemanly behavior towards patients and a balance between truthfulness and benevolent deception. Opinions about the necessity and the required depth of such deception varied, but most doctors sanctioned it, often with the complicity of patients and family members. It was only in the second half of the twentieth century that this became less acceptable to the profession and that patients began to take a more active role in inquiring about their treatment and prognosis (Lerner 2004; Rogers 2008; Sokol 2006).

Despite the fact that in the nineteenth and early twentieth centuries there were cases in which doctors were subjected to malpractice proceedings that hinged on some aspect of consent, and despite the fact that patients were informed to some extent about risks and did negotiate or decline their doctors' recommendation to undergo treatment, scholars have argued that patient autonomy and informed

consent are concepts that need to be historicized, since they acquired different meanings over time. For instance, when in 1822 the US army doctor William Beaumont retained the services of a man to study the process of digestion though an opening in his stomach that had been caused by an injury, the papers that were signed outlining the conditions under which the study was to proceed and the monetary compensation were more akin to an employment contract than to a document of informed consent (Lederer 2008; Pernick 1982; Powderly 2000).

At the turn of the twentieth century, Walter Reed's experiment in Cuba to understand the etiology of yellow fever has been seen by historians as one of the earliest precursors to our current understanding of informed consent. The researchers asked their subjects to read and sign a document outlining the potential risks, and they allowed the subjects to withdraw from the experiment at any time (Lederer 2008). In terms of legal precedent, basic informed consent in the United States has been traced to a handful of battery cases against doctors at the beginning of the twentieth century. These cases – the most cited of which is *Schloendorff v. Society of New York Hospitals* (1914) – anticipated later concepts of consent by addressing broader principles underlying the doctor-patient relationship such as the moral principle of respect for the patient's autonomy and self-determination. Nevertheless, scholars have noted that these early cases do not mean that a broad theory of consent became available immediately: legal theory develops incrementally, and in the common law system cases build upon each other further clarifying the law. It was only in 1957, a decade after the issuing of the Nuremberg Code, that the term "informed consent" was used in court in the case of *Salgo v. Leland Stanford Jr. University Board of Trustees* in which the judge ruled that doctors had a duty to disclose any facts that are necessary for the patient to give consent (Faden et al. 1986).

The Nuremberg Medical Trial, which concluded in the summer of 1947, was an unprecedented step in the effort to establish external guidelines addressing the ethical dimensions of clinical research. Although medical scientists, such as the American physiologists John West Thompson and Andrew Conway Ivy and the neurologist Leo Alexander, played a critical role in the investigation of war crimes and in the drafting of guidelines, the Nuremberg Code was external to the medical profession, being part of the verdict rendered by the tribunal in the trial of Nazi doctors who had engaged in biomedical experiments during WWII. A set of ten principles, the Code established the vital importance of voluntary consent and outlined sets of criteria for the conduct of medical experiments (Faden et al. 1986; Rothman 1996; Weindling 2001, 2004). Taking some inspiration from these principles, and after a long debate, the World Medical Association adopted in 1964 the Declaration of Helsinki as a set of safeguards for clinical research, although a glaring departure from the Nuremberg Code consisted in the fact that experimentation on those who could not give consent, such as children and prisoners, was actually sanctioned. Historians have suggested that this development reflected a strong American influence, which sought to protect the interests of pharmaceutical companies whose research on drugs and vaccines relied heavily on prison and reformatory populations (Lederer 2007).

In the immediate aftermath of the war, the AMA had also drafted Principles of Medical Ethics that required voluntary consent, prior testing on animals, and proper medical protection during the experiment. Nevertheless, despite the existence of these guidelines, many postwar researchers did not employ informed consent in their clinical studies and deliberately caused harm to their patients. The Tuskegee syphilis experiment (in which treatment for syphilis was withheld from poor African American patients in order to observe the course of the disease), the Guatemala prison experiment (in which inmates were deliberately infected with syphilis, gonorrhea, and chancroid), and the Willowbrook State School experiments (in which mentally disabled children were infected with strains of hepatitis) are among the most infamous examples of ethically egregious practices in the postwar era (Eckart 2006; Reverby 2009, 2011).

Scholars have argued that the critical period in which change occurred in the United States was between 1966 and 1976. The National Institutes of Health (NIH), which after the war had become the most important source of funding for biomedical research in the country, internally addressed potential ethical and legal concerns related to human experimentation, especially in light of the fact that experiments were increasingly using healthy subjects. The NIH adopted an internal practice of evaluating each research proposal with the help of a committee not involved in the project. Some historians have argued that this reliance on expert committees – which constituted in effect precursors to the Institutional Review Board (IRB) – reinforced professional and institutional autonomy (Stark 2012). In 1966, the NIH issued two cornerstone statements of policy that mandated signed informed consent and required proposed studies with human subjects to be reviewed by committee. At around the same time, following a doctor's whistle-blowing article about egregious practices in experimental research on human subjects and in a time period in which the rights movement was galvanizing American society, questions about human experimentation and medical decision-making were thrust into the public spotlight. In 1974 Congress passed the National Research Act, which created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Commission studied wide-ranging issues such as the ethics of experimentation on prisoners, sterilization of the mentally disabled, and psychosurgery. The question of consent was explored, and the Commission concluded that informed consent had to be based on three principles: information, comprehension, and voluntariness. The role and functioning of IRBs were also investigated, and the Commission recommended guidelines that were enacted as federal regulations (Faden et al. 1986; Rothman 1991).

The Development of Neurological Surgery

The history of modern neurosurgery began to take shape in the latter part of the nineteenth century, and the ethical dimensions of this story closely parallel the developments outlined above. Trephining (i.e., drilling a hole in the skull) was a very old procedure and had been employed to different ends throughout history,

but modern surgeons tended to open the skull only in desperate cases of skull fracture caused by severe trauma, and usually an effort was made not to incise the membrane that covers the brain. A few developments in the nineteenth century made the procedure of opening the skull less dangerous. The discovery of anesthesia just before the middle of the century, followed shortly by antiseptic and eventually aseptic practices that reduced the risk of infection, gave surgeons greater latitude in the types of operations they could perform. During the same period, Paul Broca, Gustav Fritsch, Eduard Hitzig, John Hughlings Jackson, and others began to lay the experimental and theoretical foundations of localization, realizing that discrete functions (sensory, motor, language, etc.) were localized in discrete areas of the brain. In the context of these developments, a few European surgeons – William Macewen in Glasgow, Victor Horsley in London, and Ernst von Bergmann in Berlin – started to engage in a consistent attempt to open the skull of patients who suffered from neurological symptoms indicating the presence of tumors, abscesses, or blood clots (Dagi et al. 1997).

Ethical concerns about interference with the human brain were raised quite early on. One of the most criticized experiments of the last quarter of the nineteenth century, denounced by antivivisectionists and the medical profession alike, was Roberts Bartholow's investigation of the electrical excitability of the human brain at the Good Samaritan Hospital in Cincinnati, Ohio. In 1874, Bartholow admitted to his care a "feebleminded" young woman who suffered from a rodent ulcer that had left a gaping hole in her skull. The physician, who had been aware of the recent work of European experimenters on the localization of function, obtained the woman's assent to electrically stimulate her exposed brain by inserting electrodes at various points into the brain matter. He observed muscle contraction in different parts of her body at low intensities, but when he increased the intensity of the current, he caused the woman great distress, pain, and convulsions followed by coma. She died a few days later, and following Bartholow's published report of the case, a chorus of critics, including the American Medical Association, condemned not only the experiment but also, given the patient's mental abilities, Bartholow's claim that the patient had consented (Bartholow 1874; Lederer 1995; Morgan 1982).

Generally speaking, in the last decades of the nineteenth century, the surgical community embraced an interventionist ethos, and there were frequent denunciations – both from outside and inside the profession – of the wide prevalence of unnecessary procedures. Many surgeons had felt buoyed by the arrival of antiseptic and later aseptic measures, which could, at least in principle, dramatically decrease operative mortality rates. This interventionist ethos extended to procedures on the skull and brain (Scarff 1955). Not only were many surgeons eager to attempt the removal of a suspected tumor, but controversial new therapeutic procedures were also devised, as in the case of microcephaly – a diagnosis that was applied to babies and children who were thought to have small heads, who were considered to be "idiots" or "feebleminded," or who suffered from early ossification of the fontanels. The procedure, prevalent in Europe and North America, consisted of long incisions in or resections of these patients' skull on the theory that the brain

did not have room to grow or that pressure on the brain was responsible for some of the symptoms. In 1894 the New York doctor Abraham Jacobi, in an impassioned plea against overtreatment in medicine, recounted how over the past decade "the brains of operative surgeons were taken with the *furor operandi* on the brains of luckless children" (Jacobi 1894). He presented data that suggested the procedure's high mortality rate and denounced it on the principle that a doctor's first duty was to do no harm (Feinsod and Davis 2003).

A similar – and similarly controversial – procedure was tried on patients suffering from general paralysis of the insane. In order to relieve a supposed mechanical pressure compressing the brain, the surgeons permanently removed sections of the skull. In a debate in the *British Medical Journal*, critics seemed to balk at the severity of the operation and to doubt its utility, concluding that it was not justified (Berrios 1991, 1997; Shaw 1891; Tuke 1890). Questions of consent, or the patients' ability to give consent, were not raised, just as they had not been raised by Jacobi, suggesting that even for critics the most important issue at stake concerned medical judgment, rather than a patient's (or patient's relatives') ability to make an informed decision about such new and experimental procedures. The same was true of another late nineteenth-century operation devised by the Swiss psychiatrist Gottlieb Burckhardt, a doctor deeply committed to organicist explanations of mental illness who lesioned the cortex of patients exhibiting symptoms such as delusions, hallucinations, violence, and melancholia. One patient died, but Burckhardt claimed that three out of the remaining five subjects showed improvement. The medical community did not react positively to Burckhardt's procedure, and he stopped performing it (Berrios 1997; Whitaker et al. 1996).

In the United States surgeons tried brain and skull operations for a variety of conditions, including epilepsy, traumatic insanity, intractable headache, and even for criminality. Usually, some kind of consent was sought, and surgeons sometimes elaborated on the possible dangers involved in the procedure. For instance the St. Louis, Missouri, surgeon Emory Lanphear tried surgery in a case of "softening of the brain" and claimed that "I explained to [the patient] and to his doctors that no one besides myself had ever made the proposal to operate for cerebral softening, pure and simple, and that it was purely experimental, as well as not wholly without danger. The reply was that any danger would be encountered for a bare possibility of even incomplete relief from the fearful maddening sensations in his head" (Lanphear 1895). The patient, who had had a stroke several years previously, suffered from some paralysis, aphasia, headaches, and his intellect had been "somewhat affected." It is not clear, nor does Lanphear explain, to what extent the patient's consent could be taken at face value. Similarly unquestioned was the consent of criminals who were given the choice between a surgery to relieve "pressure on the brain" and a prison sentence (Gavrus 2011b, 2013).

As brain surgery became professionalized, a certain conservatism came to initially dominate the field. In the first two decades of the twentieth century, the influential American surgeon Harvey Cushing helped establish neurosurgery as a medical specialty by refining neurosurgical knowledge and technique while also founding a specialist society and creating an international school of neurosurgery by training

fellows from all over the world. Cushing favored a conservative approach that consisted in little damage to healthy brain tissue, minimal loss of blood, and intervention in only a limited number conditions, chiefly tumors, trauma, and trigeminal neuralgia (Bliss 2005; Gavrus 2011a; Greenblatt 2003; Greenblatt and Smith 1997).

Numerous letters sent by patients suggest that Cushing was generally revered and adored by those he treated (Bliss 2005). Doctors enjoyed a great deal of prestige and public trust in the first half of the twentieth century, a reality that began to decline in the 1960s and has continued to erode ever since (Burnham 1982; Imber 2008). The authority Cushing commanded as an elite surgeon likely gave his recommendation of treatment extraordinary weight. He did seek (unwritten) consent for the surgery, but it is difficult to assess the complexity of the conversation he might have had with the patients. More challenging for Cushing appeared to have been the process of obtaining the consent of the family for postmortem examinations. Such was the 1927 case of a young woman diagnosed with an occipital meningioma. After the surgery, once it became clear that her case was hopeless, one surgeon who worked with Cushing approached the patient's husband and asked for permission to perform an autopsy after her death. The surgeon noted in the patient's chart that "[the husband] is a very decent fellow but ruled by his feeling and my best efforts were wasted on him. His one definite grievance was that he had brought his wife to the hospital for observation with the distinct understanding that he was to talk to Dr. Cushing before any operation was performed and this was not accomplished. He said he would be glad to pay [the] additional bill that the hospital would send him but could under no circumstances consent to an autopsy and finally I was forced to let the matter drop" (Cohen-Gadol and Spencer 2007).

This difficulty in obtaining consent for autopsies was longstanding, and it appears that at least in the earlier period, Cushing may have sometimes resorted to unorthodox measures. In 1911, William Sharpe, who was at the time an assistant resident in surgery at Johns Hopkins, was sent by Cushing to perform an autopsy and remove a number of organs from a former acromegaly patient. Since the family had not given consent for the procedure, Cushing had bribed the priest instead, and Sharpe performed the autopsy stealthily, at the funeral parlor, the dead body already dressed for burial and laid in the coffin. Sharpe wrote that normally "[w]henever death occurred, every possible effort was made by Dr. Cushing to obtain the nearest relative's permission for a postmortem examination – not only to ascertain the cause of death but to confirm or disprove the preoperative diagnosis. Even without permission, it was still possible to "inspect" through the operative incision or, if need be, via the rectum." Sharpe felt that Cushing was often forced to spend too much time obtaining the relatives' consent after the patient died. A few years later, when Sharpe started his own practice in New York City, he instituted a system whereby the relative signed a consent form for the postmortem before the surgery was performed, in case the patient died under the knife. "In only three cases since that time has my request for this written permission been refused," Sharpe noted, "and in each of those three cases I therefore refused to operate" (Sharpe 1952).

Cushing's somewhat conservative attitude to brain surgery was supplanted by the much more aggressive and radical ethos of the second and third generations of

neurosurgeons. Sharpe himself operated on the brains of spastic paralysis patients. Starting in the 1930s, the younger neurosurgeons devised bolder, more experimental, and more extensive surgical procedures – such as the hemispherectomy (the removal of one cerebral hemisphere in some cases of epilepsy or tumor), the sectioning of the corpus callosum, and the bilateral removal of the medial temporal lobes (Scoville and Milner 1957). This last procedure was famously employed in the case of H.M., a patient who postsurgery developed profound anterograde amnesia and became a subject in memory research over the following decades. Neurosurgeons and neurologists also devised risky and painful experimental protocols for conditions such as epilepsy, and they treated institutionalized patients in a much more cavalier manner than they did extramural patients (Dwyer 2012). The surgeons enlarged their therapeutic repertoire considerably. They were now operating extensively in cases of epilepsy (the Montreal Neurological Institute, under Wilder Penfield, becoming the most famous center for such surgeries), hydrocephalus, hypertension, Parkinson's Disease, various pain disorders, Reynauld's Disease, and psychiatric conditions (Gavrus 2011b). The lobotomy would become the most contested of these new and aggressive procedures.

The Psychosurgery Challenge

In 1935, the Portuguese neurologist Egas Moniz, who was aware of experimental reports linking frontal lobe damage to a decrease in aggression and other emotional changes, devised a procedure he called prefrontal leukotomy. Through two holes drilled into the skull, an instrument was introduced into the brain and rotated so that it cut nerve fibers connecting the frontal lobe to other areas of the brain. Moniz tried the procedure on a number of psychiatric patients and claimed that many had been cured or were substantially improved. In the United States, many second- and third-generation neurosurgeons embraced the procedure, having been personally encouraged to do so by the Yale physiologist John Fulton. The neurologist Walter Freeman, in particular, working at first with the neurosurgeon James Watts, zealously promoted lobotomy especially in cases of severe affective disorders. Freeman later devised the transorbital lobotomy, a more easily performed procedure in which the frontal lobe was reached though the eye socket. Tens of thousands of Americans were subjected to lobotomy, most of them before the mid 1950s, when alternatives such as the antipsychotic chlorpromazine became available and when reports identifying serious side effects cooled lobotomy's appeal. Its initial popularity was partly the result of a lack of alternative therapies for psychiatric patients hospitalized in overcrowded institutions. Since lobotomy was initially presented as an embodiment of modern scientific medicine, psychiatrists in state institutions found it appealing, while patients and their families invested a lot of hope in its promise to cure mental illness. The Louisville neurosurgeon Glen Spurling, for example, wrote to Fulton in 1937 that he was "besieged by families of psychopathic patients to do this wonderful operation" (Spurling to Fulton, January 5, 1937, John Fulton Correspondence, Yale University

Archives). Even psychodynamically oriented psychiatrists accepted the procedure and incorporated it into their practice (Pressman 1998; Raz 2008; Valenstein 1986, 1997).

Although a few critics had expressed opposition since its inception, lobotomy became seriously contested only in the 1970s, after it had long been supplanted by procedures touted to be more firmly anchored in anatomical and physiological knowledge – smaller, more localized lesions to healthy brain tissue, such as the cingulotomy in which fibers passing though the anterior cingulated gyrus were destroyed in severe cases of obsessive-compulsive disorder. However, when suggestions were made about the utility of such brain lesions in curbing violent tendencies, they were met by a rising public concern about social control and lack of patient autonomy. In the wake of an unprecedented debate that involved the government, professional organizations, and the public, the practice of psychosurgery underwent great changes. In 1973, the state of Oregon enacted legislation requiring the operation to be first approved by a Psychosurgery Review Board whose members determined if the treatment had clinical merit and if the patient or guardian had given voluntary and informed consent. California enacted even stricter laws mandating that the patient had to have the capacity for informed consent herself; no guardian could make that decision for the patient. In this increasingly charged legal environment, in 1977 the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research commissioned studies to verify the utility of psychosurgery and issued a report (1977) that did not recommend the banning of psychosurgery as some critics had hoped, but advised strict safeguards and regulations for its practice. Among other issues, the 1970s debate on psychosurgery raised ethical concerns about the irreversibility of the procedure, informed consent (especially in the case of involuntarily confined individuals), the constitutionality of review boards, the protection of patients, and the problem of using prisoners in research (Valenstein 1980a, b).

Conclusions: Ethics and Contemporary Neurosurgery

Following the sweeping changes that were occurring in the 1960s and 1970s, the leading professional neurosurgical organization in North America, the American Association of Neurological Surgeons, convened a committee for the first time to study and issue a code of ethics. The Neurosurgical Code emphasized beneficence, the autonomy of the individual, and justice. The president of the association expressed his hope that "the Code [...] can help us unravel the moral dilemmas that we face in our practice today, and which I fear will become increasingly numerous and complex" (Patterson 1986). Indeed, new techniques and experimental procedures such as deep brain stimulation, which was approved by the FDA in the 1990s for a variety of neurological and psychiatric conditions, continued to bring with them challenges to the ethical underpinnings of neurosurgical practice (Skuban et al. 2011). Similarly, the rapid development of functional neurosurgery – the branch of neurosurgery that uses neuroimaging techniques to target areas for ablation, stimulation, or other interventions – has raised questions about the ethics of

interference and the fine line between treatment and experiment. Most recently, debates about the ethics of brain function enhancement and brain-machine interface are drawing attention (Lipsman and Bernstein 2011). Voluntary informed consent continues to be debated in the professional literature, not only by bioethicists, but by neurosurgeons as well. Efforts continue to be made to evaluate the process of informed consent and to devise better ways of engaging patients in the process of decision-making (Kondziolka et al. 2006).

As this brief history demonstrates, consent is a concept that needs to be historicized, since its meaning and application have changed over time. The development of our contemporary understanding of voluntary informed consent has been a slow process influenced by specific historical events and debates, both internal and external to the medical profession. Historians, bioethicists, legal scholars, and, of course, doctors continue to theorize the concept of consent and to scrutinize the new developments in medical research and treatment that ignite fresh debates about the ethics of interfering with the human brain.

Cross-References

- ▶ Awake Craniotomies: Burden or Benefit for the Patient?
- ▶ Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine
- ▶ Ethical Implications of Brain Stimulation
- ▶ Ethical Implications of Brain–Computer Interfacing
- ▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
- ▶ Ethics in Neurosurgery
- ▶ Ethics of Epilepsy Surgery
- ▶ Ethics of Functional Neurosurgery
- ▶ Historical and Ethical Perspectives of Modern Neuroimaging
- ▶ History of Neuroscience and Neuroethics: Introduction
- ▶ Neurosurgery: Past, Present, and Future
- ▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives

References

Bartholow, R. (1874). Experimental investigations into the functions of the human brain. *The American Journal of the Medical Sciences, 67*, 305–313.

Bernard, C. (1865). *Introduction a l'étude de la médecine expérimentale*. Paris: J.B. Baillière.

Berrios, G. E. (1991). Psychosurgery in Britain and elsewhere: A conceptual history. In G. E. Berrios & H. Freeman (Eds.), *150 years of British psychiatry* (pp. 180–196). London: Gaskell.

Berrios, G. E. (1997). The origins of psychosurgery: Shaw, Burckhardt and Moniz. *History of Psychiatry, 8*(29 pt 1), 61–81.

Bliss, M. (2005). *Harvey Cushing: A life in surgery*. New York: Oxford University Press.
Burnham, J. C. (1982). American medicine's golden age: What happened to it? *Science, 215*(4539), 1474–1479.
Cohen-Gadol, A. A., & Spencer, D. D. (2007). *The legacy of Harvey Cushing: profiles of patient care*. Rolling Meadows: Thieme/American Association of Neurosurgeons.
Cunningham, A., & Williams, P. (1992). *The laboratory revolution in medicine*. Cambridge: Cambridge University Press.
Dagi, T. F., Epstein, M. H., & Greenblatt, S. H. (1997). *A history of neurosurgery in its scientific and professional contexts*. Park Ridge: American Association of Neurological Surgeons.
Dwyer, E. (2012). Neurological patients as experimental subjects: Epilepsy studies in the United States. In L. S. Jacyna & S. T. Casper (Eds.), *The neurological patient in history* (pp. 44–60). Rochester: University of Rochester Press.
Eckart, W. U. (Ed.). (2006). *Man, medicine, and the state: The human body as an object of government sponsored medical research in the 20th century*. Stuttgart: Steiner.
Faden, R. R., Beauchamp, T. L., & King, N. M. P. (1986). *A history and theory of informed consent*. New York: Oxford University Press.
Feinsod, M., & Davis, N. L. (2003). Unlocking the brain: Attempts to improve mental function of microcephalic retarded children by "craniotomy". *Neurosurgery, 53*(3), 723–730.
Gavrus, D. (2011a). Men of dreams and men of action: Neurologists, neurosurgeons, and the performance of professional identity, 1920–1950. *Bulletin of the History of Medicine, 85*(1), 57–92.
Gavrus, D. (2011b). *Men of strong opinions: Identity, self-representation, and the performance of neurosurgery, 1919–1950*. (Ph.D. thesis). University of Toronto, Toronto.
Gavrus, D. (2013). *'Teaching morality with a surgeon's scalpel:' Brain surgery for criminals during the Progressive Era*. Paper presented at the American Association for the History of Medicine, Atlanta.
Greenblatt, S. H. (2003). Harvey Cushing's paradigmatic contribution to neurosurgery and the evolution of his thoughts about specialization. *Bulletin of the History of Medicine, 77*(4), 789–822.
Greenblatt, S. H., & Smith, D. (1997). The emergence of Cushing's leadership: 1901–1920. In T. F. Dagi, M. H. Epstein, & S. H. Greenblatt (Eds.), *A history of neurosurgery in its scientific and professional contexts* (pp. 167–190). Park Ridge: American Association of Neurological Surgeons.
Imber, J. B. (2008). *Trusting doctors: The decline of moral authority in American medicine*. Princeton: Princeton University Press.
Jacobi, A. (1894). Non nocere. *Medical Record, 45*(20), 609–619.
Kondziolka, D. S., Pirris, S. M., & Lunsford, L. D. (2006). Improving the informed consent process for surgery. *Neurosurgery, 58*(6).
Lanphear, E. (1895). Lectures on intracranial surgery; XI. The surgical treatment of insanity. *Journal of the American Medical Association, 24*(23), 883–886.
Lederer, S. E. (1995). *Subjected to science: Human experimentation in America before the Second World War*. Baltimore: Johns Hopkins University Press.
Lederer, S. E. (2007). Research without borders: The origins of the declaration of Helsinki. In U. Schmidt & A. Frewer (Eds.), *History and theory of human experimentation: The declaration of Helsinki and modern medical ethics* (pp. 145–164). Stuttgart: Steiner.
Lederer, S. E. (2008). Walter Reed and the yellow fever experiments. In E. J. Emanuel, C. C. Grady, R. A. Crouch, R. K. Lie, F. G. Miller, & D. D. Wendler (Eds.), *The Oxford textbook of clinical research ethics* (pp. 9–17). Oxford: Oxford University Press.
Lerner, B. H. (2004). Beyond informed consent: Did cancer patients challenge their physicians in the post-World War II era? *Journal of the History of Medicine and Allied Sciences, 59*(4), 507–521.
Lipsman, N., & Bernstein, M. (2011). Ethical issues in functional neurosurgery: Emerging applications and controversies. In J. Illes & B. J. Sahakian (Eds.), *Oxford handbook of neuroethics* (pp. 405–416). Oxford: Oxford University Press.
Morgan, J. P. (1982). The first reported case of electrical stimulation of the human brain. *Journal of the History of Medicine and Allied Sciences, 37*(1), 51–64.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1977). *Report and recommendations on psychosurgery*, Bethesda.

Patterson, R. H. (1986). A code of ethics: The 1986 AANS presidential address. *Journal of Neurosurgery, 65*(9), 271–277.

Pernick, M. S. (1982). The Patient's role in medical decision-making: A social history of informed consent in medical therapy. In *President's commission for the study of ethical problems in medicine and biomedical and behavioral research, Making health care decisions* (Vol. 3, pp. 1–35). Washington, DC: U.S. Government Printing Office.

Powderly, K. E. (2000). Patient consent and negotiation in the Brooklyn gynecological practice of Alexander J.C. Skene: 1863–1900. *The Journal of Medicine and Philosophy, 25*(1), 12–27.

Pressman, J. (1998). *Last resort: Psychosurgery and the limits of medicine*. Cambridge, UK: Cambridge University Press.

Raz, M. (2008). Between the ego and the icepick: Psychosurgery, psychoanalysis, and psychiatric discourse. *Bulletin of the History of Medicine, 82*(2), 387–420.

Reverby, S. M. (2009). *Examining Tuskegee: The infamous syphilis study and its legacy*. Chapel Hill: University of North Carolina Press.

Reverby, S. M. (2011). "Normal Exposure" and inoculation syphilis: A PHS "Tuskegee" doctor in Guatemala, 1946–1948. *Journal of Policy History, 23*(1), 6–28.

Rogers, N. (2008). 'Silence has its own Stories': Elizabeth Kenny, polio and the culture of medicine. *Social History of Medicine, 21*(1), 145–161.

Rothman, D. J. (1991). *Strangers at the bedside: A history of how law and bioethics transformed medical decision making*. New York: Basic Books.

Rothman, D. J. (1996). *Other people's bodies: Human experimentation on the 50th anniversary of the Nuremberg Code*. Paper presented at the American Osler Society, John P. McGovern Award Lectureship, San Francisco.

Scarff, J. E. (1955). Fifty years of neurosurgery, 1905–1955. *International Abstracts of Surgery, 101*(5), 417–513.

Schlich, T. (2007). Surgery, science and modernity: Operating rooms and laboratories as spaces of control. *History of Science, 45*(149), 231–256.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20*(1), 11–21.

Sharpe, W. (1952). *Brain surgeon: The autobiography of William Sharpe*. New York: Viking Press.

Shaw, C. T. (1891). Surgical treatment of general paralysis of the insane. *British Medical Journal, 2*(1602), 581–583.

Shortt, S. E. (1983). Physicians, science, and status: Issues in the professionalization of Anglo-American medicine in the nineteenth century. *Medical History, 27*(1), 51–68.

Skuban, T., Hardenacke, K., Woopen, C., & Kuhn, J. (2011). Informed consent in deep brain stimulation – Ethical considerations in a stress field of pride and prejudice. *Frontiers in Integrative Neuroscience, 5*, 7.

Sokol, D. K. (2006). How the doctor's nose has shortened over time; a historical overview of the truth-telling debate in the doctor-patient relationship. *Journal of the Royal Society of Medicine, 99*(12), 632–636.

Stark, L. (2012). *Behind closed doors: IRBs and the making of ethical research*. Chicago: University of Chicago Press.

Trohler, U. (2007). The long road of moral concern: Doctors' ethos and statute law relating to human research in Europe. In U. Schmidt & A. Frewer (Eds.), *History and theory of human experimentation: The declaration of Helsinki and modern medical ethics*. Stuttgart: Steiner.

Tuke, J. B. (1890). The surgical treatment of intracranial fluid pressure. *British Medical Journal, 1*(1514), 8–11.

Valenstein, E. S. (1980a). Historical perspective. In E. S. Valenstein (Ed.), *The psychosurgery debate: Scientific, legal, and ethical perspectives*. San Francisco: W. H. Freeman.

Valenstein, E. S. (Ed.). (1980b). *The psychosurgery debate: Scientific, legal, and ethical perspectives*. San Francisco: W. H. Freeman.

Valenstein, E. S. (1986). *Great and desperate cures: The rise and decline of psychosurgery and other radical treatments for mental illness*. New York: Basic Books.

Valenstein, E. S. (1997). History of psychosurgery. In T. F. Dagi, M. H. Epstein, & S. H. Greenblatt (Eds.), *A history of neurosurgery in its scientific and professional contexts* (pp. 499–516). Park Ridge: American Association of Neurological Surgeons.

Warner, J. H. (1986). *The therapeutic perspective: Medical practice, knowledge, and identity in America, 1820–1885*. Cambridge, MA: Harvard University Press.

Warner, J. H. (1991). Ideals of science and their discontents in late nineteenth-century American medicine. *Isis, 82*(313), 454–478.

Weindling, P. (2001). The origins of informed consent: The international scientific commission on medical war crimes, and the Nuremburg code. *Bulletin of the History of Medicine, 75*(1), 37–71.

Weindling, P. (2004). *Nazi medicine and the Nuremberg trials: From medical war crimes to informed consent*. New York: Palgrave Macmillan.

Whitaker, H. A., Stemmer, B., & Joanette, Y. (1996). A psychosurgical chapter in the history of cerebral localization: The six cases of Gottlieb Burckhardt. In C. Code, C. W. Wallesch, A. R. Lecours, & Y. Joanette (Eds.), *Classic cases in neuropsychology* (pp. 275–304). London: Erlbaum.

Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry

32

Frank W. Stahnisch

Contents

Abstract

This chapter discusses some early developments in patient-centered treatments in nineteenth-century psychiatry as well as the backlash towards physician-oriented and frequently oppressive treatment and research styles in biological psychiatry at the beginning of the twentieth century. The chapter also aims at tracing some of the modern neuromanipulative approaches of deep brain stimulation back to their historical origins and comparing them with more outmoded approaches of the electrophysiological alteration of the human cortex and deep brain structures. In the second part of this chapter, an overview on the development of modern deep brain stimulation methods and related ethical problem fields is provided. This chapter adds another perspective to the neuroethical discussion by putting forward history of medicine and neuroscience case examples regarding the ethical problems involved, focusing on the issue of medical manipulation during the technological development of modern biological

F.W. Stahnisch
Department of Community Health Sciences and Department of History, Hotchkiss Brain Institute/ Institute for Public Health, The University of Calgary, Calgary, AB, Canada
e-mail: fwstahni@ucalgary.ca; frank.stahnisch@berkeley.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_25,

psychiatry. It is shown that, following to the Second World War, electrophysiological stimulation and shock approaches were developed, which began to crucially change the treatment options for mentally ill patients. By means of a comparative analysis, it is argued that many contemporary debates which question neuroethical applications are problematic in significant respects. Of major concern, in this regard, is the often blurred conceptual boundary that is furnished by the complex relationships between clinical research, therapeutic intentions, and the perspective on patient autonomy.

Introduction

Until the time of the French physician Philippe Pinel (1745–1826), who was instrumental in the process of unchaining the insane at the Bicêtre mental institution in Paris at the end of the eighteenth century, diagnostic and therapeutic options for patients with mental illness had been very poor and limited to contingent forms of care by the family, the local community, and legal guardians acting from philanthropic and charity perspectives (Ackerknecht 1968). With the advent of larger-scale hospitals and mental asylums during the middle of the nineteenth century and more specialized research approaches towards the infirmities and demands of the mentally ill, the repertoire of diagnostic techniques and therapeutic options increased largely towards the *Fin de Siècle* (Reiser 1978). Certainly, the development of nonrestraint measures of custodianship and caring ("freeing the mentally ill from their chains"), the introduction of physical therapies (water cures and electro-puncture therapy), as well as "modern" shock therapies (malaria fever cures, insulin, and electroconvulsive shock therapies) at the beginning of the twentieth century not only brought with them rising social concerns about the innovative practices in psychiatry but also led increasingly to redefinitions of medical ethics and the context in which these new diagnostic and therapeutic technologies became applied (Leahey 2000) (► Chap. 70, "Human Brain Research and Ethics").

The blessings of modernized treatment approaches of the mentally and nervously ill became themselves a major subject of ethical debates in psychiatry and neuroethics. Early and notorious shock therapies of the twentieth century have morphed into a paradigmatic example of the limits of medicine, showing how the best intentions of doctors and researchers can go ever so wrong, turning often into a curse for those already suffering in mental institutions (Pressman 1998). Social concerns about ethical behaviors and the doctor-patient relationship in psychiatry and related fields were frequently born out of the introduction of new healing technologies and institutionalized hospital care. Most bioethical discourses in this field – whether related to physical and shock therapies, the early use of deep brain stimulation (► Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"; ► Chap. 39, "Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review") in psychiatric contexts, or the later introduction of psychoactive drugs and pharmaceutical treatments

(Shorter 2009) – centered around pivotal issues regarding external influences on test persons in research settings or on patients' autonomy in clinical hospital contexts (Lederer 1995; Roelcke 2004), as shall be further examined in the following sections of this chapter.

Nonrestraint Therapies Since the End of the Eighteenth Century

The European Enlightenment of the eighteenth century gave rise to new conceptualizations of "human rationality," "personhood," and the "mind-brain relationship" that were not simply of philosophical interest but integrated with dramatically changing social attitudes towards the lower classes, women, and children, as well as the mentally ill (Rousseau 1990) (▶ Chap. 58, "Relationship of Benefits to Risks in Psychiatric Research Interventions"). Contemporary forms of treatment for the mentally ill began to center on "moral therapeutic" approaches, which understood the deranged human mind as an effect of the often oppressive and harmful social conditions (Ackerknecht 1986). During the postrevolutionary period in French medicine, treatment of the mentally ill transitioned from traditional remedies in Christian and aristocratic social settings – e.g., bloodletting, emetics, and purging – towards political and medical therapies which had been promoted by Pinel and other reformers as the new moral treatment in psychiatry. Hospitals became now the predominant place for attempting to treat psychiatric patients, rather than custodial institutions in which the patients had been traditionally housed for years and even decades (Lindemann 1999). At the same time, more specialized medical education programs for mental and nervous illness were integrated in the hospitals; the predominant teaching style in medicine and psychiatry became practical instruction at the bedside vis-à-vis the patient (Weisz 1994).

These developments had social implications for the physician-patient relationship, which was vivibly transformed in the "hospital medicine context" of the early nineteenth century. The diagnostic process now moved away from being a rather mutual venture between the doctor and the patient, as the introduction of new forms of technological instruments, such as the stethoscope, the thermometer, and later the reflex hammer, began to create a physical and mental distance between doctors and their patients (Porter 1995). Physicians – and not soon after, the members of the new profession of psychiatry – began to standardize moral treatments for all patients and thereby further reduce the role of individuality and personality (Stahnisch 2012). The communication between patient and physician became increasingly minimal, as the patients were rendered primarily into objects for study ("patient material" in the language of the time) rather than as equal partners of the physician.

Nineteenth-century advancements in experimental physiology and anatomical pathology translated relatively quickly into significant changes in the field of neurological diagnostics and therapy (Brazier 1988), even while the clinical treatment options in the field of psychiatry or asylum care remained very much the same (Porter 2001). Particularly academic professors of psychiatry – in France and Germany – criticized the dominant role that asylums had assumed in psychiatry, as

they regarded them to be unacceptably overcrowded with their medical staff clinging to inappropriate forms of mechanical restraint (Engstrom 2004). Psychiatric reformers saw the contemporary asylums as lacking adequate ways of patient observation at early stages of mental and nervous disorders, while clinical research needed to become more rigorous and systematic. If one takes a closer look at de- and regenerative research endeavors, for example, then it becomes evident that broader social concepts infiltrated both basic research and clinical practice at the turn of the century: One of the most influential directions of such a knowledge transfer was the psychiatric concept of "neuropathy," which the Berlin doyen of brain psychiatry Wilhelm Griesinger (1817–1868) had introduced as a reinterpretation of psychiatric brain diseases (Engstrom 2004).

Together with German internist Carl Reinhold August Wunderlich (1815–1877) and the surgeon Wilhelm Roser (1817–1888), Griesinger held many programmatic concepts of modern medical practice as being based on the natural sciences. In his medical practice at the mental asylum of Winnenthal, where he wrote his famous handbook on "*Pathology and Therapy of the Illness of the Mind* " (1845), he came to side with the earlier French "nonrestraint movement" (Griesinger 1845). As Griesinger later saw it, research in psychiatry had rather stagnated still during the 1860s and revolved around abstract theories. These limitations should drive young scientists – such as the famous Sigmund Freud (1856–1939) from Vienna – to study hypnosis as an innovative therapeutic option for neurology under Jean-Martin Charcot (1825–1893) in Paris. Renewed interest in clinical research also led to a reemphasis in the use of quantification approaches, hospital graphs and charts, and the innovative creation of instruments, such as the recent electro-puncture devices, which Griesinger integrated into psychiatry (Stahnisch 2007).

When the social reformer Griesinger accepted a position at the Charité in Berlin in 1865, his patients were housed either in the wards of the medical clinic or, together with the syphilitics and sick prisoners (Temkin 1955), in a closed asylum. However, the board of directors offered Griesinger a neurological ward with the outpatient department at his disposal, and he soon separated the psychiatric from the neurological patients. He actively promoted his new approaches to psychiatric care, such as implementing hospital care for acute patients in the middle of the capital city, while establishing an "agrarian colony" for chronic patients in the suburban outskirts as forms of community-oriented psychiatry care (Sammet 1997).

The doyen of twentieth-century psychiatry, Munich professor Emil Kraepelin (1856–1926), was markedly influenced by Griesinger, while also criticizing many of the methodological steps that Griesinger had taken with the program of "brain psychiatry." For Emil Kraepelin, as for later followers of Griesinger, the pathogenic element of "irritable weakness" was the basic constitutional concept for nervous diseases (Hagner 2000). Kraepelin alluded here to French-Swiss Bénédict Adolphe Morel's (1809–1873) concept of nervous "degeneration," which the latter had published in his "Treatise on Physical, Intellectual and Moral Degeneration of the Human Race" (1857–1858). Brain psychiatrists' perspectives on hereditary "neuropathic dispositions" also reflected recent developments in sociocultural contexts of urbanization, industrialization, and the medical implications of the labor question. These issues were

now discussed by prominent psychiatrists and neurologists, such as Kraepelin, Hugo Wilhelm von Ziemssen (1829–1902), or Carl Westphal (1833–1902), focusing on "nervous degeneration" as a rhetorical device (Schmiedebach 2001).

Kraepelin introduced the concept of surveillance wards in the late 1880s, which became a major foundation of Germany's scientific preeminence in psychiatry by the mid-1890s. Patient files and diagnostic cards were introduced as essential tools of Kraepelin's nosology and as a reflection of his emphasis on epidemiological longitudinal analyses. With the support of the US Rockefeller Foundation, the German Research Institute for Psychiatry emerged as the most eminent research institute in the 1920s and 1930s. It also created the support basis for multiple editions of Kraepelin's classic "Handbook of Psychiatry" (1915) that became the founding basis of the "Diagnostic and Statistical Manual of Mental Disorders" (DSM) in Psychiatry (Engstrom and Weber 2007). These developments were typical for the unique situation of the German-speaking medical community, where neurology and psychiatry stayed united much longer than in the Anglo-Saxon and French contexts (Karenberg 2007). Attention should be drawn also to the particularities of modern psychosomatics, as it developed as a response to the medical reductionisms in the late nineteenth and early twentieth century, often associated with Swiss-German physician Georg Groddeck (1866–1934) (Danzer 1992). He developed a holistic approach integrated with elements of deep psychology, neurology, narrative literature, and physical therapy that avoided the limitations of nineteenth-century organ-centered medicine. Groddeck became very influential on later neurologists, such as Viktor von Weizsaecker (1886–1957), who is a particularly good example for his integration of philosophy, social psychiatry, and neuroscientific innovations into the "holistic neurology" of the Weimar Period (Harrington 1996). Just as the Weimar Republic generally aimed at a reunification and healing of its fragmented society, Viktor von Weizsaecker thought to transcend the individualized perspectives on the body and mind by applying social dimensions to medical therapeutics in his widely received work, "Social Disease and Social Healing," which emphasized medical reconstruction endeavors (new forms of rehabilitation, technological innovations in medical prosthetics, and long-term care options for the chronically ill veterans) (Weizsaecker 1930) (▶ Chap. 58, "Relationship of Benefits to Risks in Psychiatric Research Interventions").

The Development of Physical and Shock Therapies Since the End of the Nineteenth Century

The impact of the new physical and shock therapies since the end of the nineteenth century cannot be overstated from an ethics perspective. The important effects of international relationships in psychiatry and the clinical neurosciences along with the exodus of Jewish and oppositional neuroscientists from German-speaking countries in the 1930s also need to be emphasized. Many psychoanalysts, such as the Vienna-trained founder of modern psychosomatics Franz Gabriel Alexander (1891–1964), Frankfurt neurologist Leo Alexander (1905–1985), Sandor Radó (1890–1972) from Budapest, Helene Deutsch (1884–1982) from Vienna, or the German émigré

psychiatrist Charles Fischer (1902–1987), shaped North American psychiatry, mental health, and psychology in dramatic ways (Grob 1983). With regard to biological psychiatry, the "emigration" of insulin shock therapy, psychiatric genetics, and neurological synapse research and their relationship with basic neuroscience in North America cannot be underestimated. The double *volte-face* from the praise of psychoanalysis over brain psychiatry, its later antagonism with genetic and biological psychiatry, and finally repulsion towards it due to the advances in molecular medicine and neuroscience was another groundbreaking process in the history of American neuroscience (Magoun 2002).

According to Jack Pressman's "Last Resort: Psychosurgery and the Limits of Medicine" (1998), ongoing reevaluations of mental health practices towards the development of aggressive interventional psychiatric therapies were already underway during the first half of the 1930s (Pressman 1998). The new somatic forms of therapy were particularly associated with Egas Moniz' (1874–1955) psychosurgery in Lisbon and Ugo Cerletti (1877–1963) and Lucio Bini (1908–1964) at La Sapienza University in Italy in 1938, where electroshock convulsion therapy (ECT) developed (Pressman 1998). ECT quickly became more popular than insulin or Metrazol therapies because less hospital staff was needed and patients did not experience as many negative side effects when becoming unconscious almost immediately. Precautions were taken to ensure the right type of patients underwent ECT; those with a history of heart problems and arthritis were often excluded as candidates. When the shock was administered, patients experienced immediate convulsions and many stopped breathing momentarily. Up to half an hour following the treatment, patients were in a comatose state, in which they could become mentally confused and quite often physically agitated. Like other somatic treatments, there were respiratory and cardiac concerns, occurrence of fractures, and some symptoms of syndromes that appeared like the original mental disease being treated but were not (Bellak 1948). Similar to Metrazol shock therapy, ECT was found to be more effective in treating depressed patients versus the schizophrenics for which it was originally intended. By 1941, the primary course of treatment for a variety of mental illnesses was ECT or sometimes a combination of ECT and insulin coma therapy (Valenstein 1948).

When the Ukraine-born and Vienna-trained neurophysiologist Manfred Sakel (1900–1957) noticed the effect of hypoglycemic states in drug addicts, where insulin had positive effects on detoxification, he thought that he had discovered a new principle of psychiatric therapy by inducing artificial coma states in schizophrenia and severe depression. The success story of insulin therapy was largely the work of eminent Swiss psychiatrist Max Mueller (1894–1980), professor and chair of psychiatry at the University of Bern, who took over Sakel's procedure in his psychiatric asylum in Muensingen as form of "active therapy" vis-à-vis former noninvasive approaches, which subsequently became widely applied in major psychiatric centers in France, Spain, and Nazi Germany (Roelcke 2005). At the time a report of the Freiburg psychiatrist Egon Kueppers (1887–1935) to the Research Ministry of the *Reich* had been emphatically received, as it bolstered the Nazi medical philosophy which saw psychiatric patients merely as "useless eaters." The shock treatments thus offered an economical alternative to

hospital-based therapy, supplementing public health measures as applied to the "organism of the people" (Borck 2005). Sakel became later himself forced into North American exile after Austria was annexed to Nazi Germany in 1938. Once he reached the USA, he started to work as a staff-attending physician at the Harlem Valley State Hospital in New York, where his forced migration from Austria led to an unintended adaptation of shock therapies in American medical communities. The problem of large hospitalization numbers had also been recognized as a major strain on the American public health care system (Grob 1983). In this situation, the new shock therapies were perceived by psychiatrists in American and Canadian psychiatric clinics and asylums as a major technological form of relief (Pressman 1998). Where eugenics and psychiatric genetics had paved the conceptual and social way for a widespread application of the new shock therapies on the old continent, the belief in technological progress and social blessings there from provided an enabling ideology on the other side of the Atlantic, leading to a groundbreaking transformation in the development of modern neuroscience in North America (see specifically in Magoun 2002) (▶ Chap. 53, "Neural Transplantation and Medical Ethics: A Contemporary History of Moral Concerns Regarding Cerebral Stem Cell and Gene Therapy").

At this time, four somatic or physical treatments had been introduced to deal with what was seen as a rising problem of mental illness. Among these were malaria and insulin coma therapy, Metrazol shock therapy, and electroconvulsive shock therapy (Valenstein 1986). Despite the hope for these particular treatments, there were several downsides to the procedures, including extended comas, irregular cardiovascular and respiratory responses, and potential damage to the brain (Bellak 1948). Although some treatment protocols were deemed "successful" in a variety of cases, additional somatic treatments continued to be tested and adopted by the medical community in hopes of more effective solutions. Another treatment was Metrazol or Cardiazol shock therapy. This method was first tested in Budapest by László (Ladislaus von) Meduna (1896–1964) in 1934. Although the initial experiments used a variety of drugs, Meduna eventually found that the injection of Metrazol quickly induced convulsions and seemed to be effective in treating dementia praecox (schizophrenia) (see also Vidal and Ortega 2011). After injection, the convulsions only took 30 seconds to begin, but during that small period of time, patients experienced a very strong and antagonizing fear of death. As many as 30 treatments were recommended or until improvement was evident. Similar to insulin shock, cardiovascular problems often occurred. Another side effect was broken bones in the spine or other parts of the skeleton due to the severity of the muscular convulsions (Bellak 1948). Although this form of treatment was quickly adopted into regular treatment regimes, by 1938 the medical community found that Metrazol shock therapy was better suited for depressed rather than schizophrenic patients. Furthermore, with the advent of ECT, the use of Metrazol slowly declined, since ECT was perceived as a safer technique for patients (Valenstein 1986). The common link between the above three somatic treatments, and also the major ethical and social concern with them, was that they had limited effectiveness and schizophrenia was still not being treated as effectively as was desired by

patients, families, and physicians, while its treatments were perceived as very dangerous, with many side effects – despite the comforting rhetoric of contemporary psychiatrists and physicians (Valenstein 1948).

Another form of invasive and drastic treatment in clinical psychiatry and neurology that should also be mentioned here – as it had been developed in its precursor stages during the first decades of the twentieth century and gained later prominence in the 1940s and 1950s – is psychosurgery. Before psychosurgery took shape in its most well-known form today (as lobotomy operations), there were many surgeons and physicians who had begun to flirt with the idea that operational, neurosurgical means might be able to ameliorate mental disease. Most prominently mentioned in the literature is the Swiss psychiatrist and mental asylum director Gottlieb Burckhardt (1836–1907), who completed the first modern psychosurgical attempt in 1888. Inspired by the laboratory animal experiments on cortical function by the Berlin neurophysiologists Gustav Theodor Fritsch (1837–1927) and Eduard Hitzig (1839–1907) (Hagner 2000), Burckhardt's procedure attempted to remove portions of the cerebral cortex of half a dozen schizophrenic patients. However his procedure, in the following, proved to be fairly unpopular among medical practitioners because of the high bleeding and infection risks associated with it. The oppositional voices in the clinical community became increasingly louder, and Burckhardt was finally ordered by the respective district college of physicians to stop his clinical and research program (Stone 2001) (▶ Chap. 31, "Informed Consent and the History of Modern Neurosurgery").

Psychosurgery – as single surgical procedure or in combination with neurophysiological stimulation measures of the brain and spinal cord (Gavrus 2011) – is a form of neuropsychiatric treatment that was and still continues to be heavily criticized from both medical practitioners and patient advocates today. The term itself can be defined in a number of ways, but a dominant interpretation holds that all surgical procedures, which seek to manipulate the brain and spinal cord tissue in order to change higher (mostly cognitive) psychological functions aligned with specific psychiatric disorders, would be part of psychosurgery. Often these clinical phenomena present themselves without any prior morphological pathologies of the brain, while psychosurgery is seen as a restitution of functional circuits, maladaptations, or microscopical lesions (Feldman et al. 2001). The most commonly used procedures were cingulotomies in the 1920s, prefrontal lobotomies in the 1930s and 1940s, and thalamo- and nerve tractomies in the 1950s and 1960s (Alesch and Keith 2004). Today, the term describes operations such as cingulotomies used to treat severe obsessive compulsive disorders, while the discredited historical notion of "psychosurgery" has become almost obsolete (Glannon 2007) (▶ Chap. 59, "Ethics in Neurosurgery").

Twentieth-Century Brain Stimulation Contexts in Psychiatry and Nervous Diseases

Shortly before and after the Second World War, deep brain stimulation approaches were developed by neurologists and psychiatrists such as

Otfrid Foerster (1873–1941) at the University of Breslau in Germany, Wilder Penfield (1891–1976) at McGill University in Montreal, or Robert Galbraith Heath (1915–1999) at the Tulane School of Medicine in New Orleans, which eventually converged into various modern methods of deep brain stimulation as they are practiced today (Baumeister 2000) (▶ Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"; ▶ Chap. 39, "Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review"). Many modern-day approaches (concepts, neurostimulatory procedures, but also ethical concerns) are based on the therapeutic and research foundations that were laid by such early neuroscientific teams. Modern neuroscientists often used very similar arguments to legitimize their research approaches to those of their predecessors, for example, in appealing to the "critical or incurable condition" of their patients or the alternative risks of other healing methods (Alesch and Keith 2004).

A more in-depth scrutiny of the early phase of brain stimulatory approaches from a historical perspective is certainly important since it demonstrates a rhetorical foundation which reaches back to the 1930s (Penfield 1978) and the early postwar period. These historical brain stimulatory developments – with their strong clinical research orientation – were finally able to lay the foundations for the current discussion of "a technological manipulation of the brain," which began raising important neuroethical questions many decades ago (Delgado 1969). The Canadian neurosurgeon Wilder Penfield learned his operative method and technique for neurostimulation principally during two extensive research periods with the neurological surgeon Otfrid Foerster at the Breslau clinical department for nervous diseases and psychiatry in 1928 and 1931 (Feindel 1998) – a fact which is little known. His German mentor had initially used the method of deep brain stimulation (stimulation of the septum of the diencephalon) for the treatment of schizophrenic patients and in individuals with posttraumatic epileptic disorders. Epileptic conditions following brain injuries were very common after the First World War, which had brought tens of thousands of patients with head wounds into neurological departments all over Europe and North America. Foerster's initial therapeutic approach would later be inversed in Penfield's program with its increasing emphasis on research into the localization of the human cerebral cortex. After returning from his tour of Europe, Penfield began successfully establishing a comparable and very modern clinical and research center at the Montreal Neurological Institute (Feindel 1992), where he closely monitored and recorded the movement effects, sensory perceptions, and psychological changes in his neurostimulated patients. During the neurosurgical operation these were kept fully conscious to avoid major damage of the brain under intraoperative stimulation, especially of the language, visual, and somatosensory centers. The electrophysiological studies on the brains of his patients in the operation theaters at the MNI built the foundation for Penfield's famous cortical cartographic map of the functional homunculus of motor and sensory cortical areas of the human brain (Finger 2000).

What is quite striking during this time, Penfield's patients (similar to those of Sakel, Cerletti, or Bini in their research on shock therapies) were not instructed

about the research intention of the general neuroscientific program, after they had agreed to receive surgical treatment to cure their conditions. This internalistic combination of neuroscientific investigations, epistemic conjectures, and (at the time paternalistic) ethical considerations did not really find its transparent introduction into the doctor-patient relationship before the 1960s. Instead ethical deliberations continued to be restricted to the individual consciousness and decision-making of the operating surgeon and some preliminary discussion with expert colleagues or hospital staff discussions which centered on the feasibility of the operations rather than on their immediate ethical concerns (Tone 2005). These paternalistic stances are omnipresent in Penfield's writings, although he barely touches on them in terms of a scholarly and critical reflection in his writings (Penfield 1977). Rather in the form of personal accounts and after discussions, such reflections can be found, as, for example, in the Proceedings of the Harvey Cushing Memorial Meetings of the Montreal Neurological Society of November 1939. Here, Penfield clearly embraced the teacher-pupil relation in clinical medicine and gives it visible priority vis-à-vis theoretical and philosophical reflections on the changes in the experimental nature of modern clinical neuroscience (Fins 2008).

In fact, not many sources can be identified that reveal Penfield's personal ethical views on medicine – and in this sense, he appears to be no exception to most of the neurosurgeons and research psychiatrists of the first half of the twentieth century. The paternalistic approach from a rather individualistic position in the ethical considerations of the Montreal neurosurgeon was quite typical for the immediate postwar period in medicine (see also Roelcke 2004). In general, Penfield was most interested in his ongoing research program and the investigation of the interrelation between the behavior-dependent and motor responses of his patients under intraoperative conditions (Penfield 1977). The primary concern of his approach, nevertheless, stayed with the surgical treatment of those areas of the brain which were seen as causally relevant in the etiology of epilepsies and tumors. It is quite remarkable, in retrospect, that although his program became so well known and Penfield's neurophysiological findings made a strong impression on the national and international scientific communities, there appears to be not much public discussion that accompanied these clinical research approaches during their advent (Fins 2008).

In contrast to Penfield's program, the public resonance which the neurostimulatory working group around Robert G. Heath at the biological psychiatry department of the Tulane School of Medicine in New Orleans garnered was enormous, developing into a central subject of media interest and provoking strong public controversy in the USA. At least in this instance, the disciplinary historiography has credited one of the co-workers of the Tulane group (Alesch and Mullet 2004), the neurosurgeon Donald E. Richardson (1919–1969), as one of the pioneers in the development of modern brain stimulation (Richardson 1980). The research in Louisiana, which commenced in the early 1950s, went undetected by the public only in its first few years. It was likewise scarcely heeded by the neuroscientific

community, probably because of the secluded nature of psychiatry and neurology communities at that time. For their ethical evaluation, it is not insignificant that nearly one hundred patients had been included in the clinical applications and the research program extended over the long time period of 30 years (Baumeister 2000). In 1996, Heath was even awarded the Gold Medal of the American Society of Biological Psychiatry, as he had "accomplished important work as a pioneer in the field of biological psychiatry" and "strongly enhanced our knowledge about the structural and functional relations of schizophrenic diseases to the brain." Over the three decades of this program, it is possible to find shifts in patient access, the doctor-patient relation, and to detect ethical reflections and rhetorical struggles within the contemporary scientific community, concerning the usefulness of Heath's procedure as being "a last desperate cure" (Fradelos 2008).

Heath – similar to Otfrid Foerster in Breslau – began his investigations within a psychiatric context: In a seventeen-year-old schizophrenic girl, Heath applied electrical currents, by means of deep brain electrodes, to the septum region of the diencephalon. Following the neurobiological understanding of the time, he wanted to "recalibrate" the functional inequality between the primitive brain regions of human development and complex cortical structures (Heath 1954). The patients were not specifically informed about the research character of these operations and most of them had psychiatric disease conditions, what for the Heath group seemed to legitimize such experiments, as the patients were unable to understand the complexity of the clinical trials and to give their autonomous consent to the physicians. The relatives and friends of these psychiatric patients became aware of the current research practices only after decades had passed, and the experiments began to attract greater media interest (Slattery 1990–1991). Not only did it seem that the patients were selected for their specific use in this research program of biological psychiatry – when explicit preference was given to severe cases in which no ethical consent could be obtained nor opposition expected – but also that the research results were being interpreted and presented in an overly positive light. It is interesting to note that in the research setting of the Tulane neurosurgeons, the electrodes for neurostimulation were not confined to the septum region. Instead, the Tulane researchers applied their electrodes to other brain regions "of interest," such as the thalamus or the nucleus caudatus, trying to find out what would happen in their patients. In so doing, these clinical researchers sought to derive further knowledge from the ad hoc intraoperative experiments, conceived when the skull of their patients lay open in front of them (Richerdson 1977).

It was mainly due to the new technological advances in medicine – more so than the sporadic forms of public media critique and patient activism – that major changes in clinical attitudes came about, a development which historical introductions in textbooks of deep brain stimulation often tend to neglect (Alesch and Mullett 2004). With the increasing use of psycho- and neuropharmacological substances, such as chlorpromazine in schizophrenia, valproate in epilepsy, and L-dopa in Parkinson's disease (▶ Chap. 30, "History of Psychopharmacology: From Functional Restitution to Functional Enhancement") brain stimulation methods lost their

appeal during the 1970s – in the United States and elsewhere (Pressman 1998). They became confined to narrow areas of functional restitution in therapy-resistant schizophrenias and intractable thalamic pain disorders.

Conclusion

The individual cases presented in this chapter expose some of the conceptual ambivalences in the complex relationship between therapeutic intentions – by biological psychiatrists and clinical neurologists – and the effects (social and personal) of medical research. Since the early days of hospitalizing the mentally ill and the shock therapies and electrostimulation programs from the 1900s to the 1930s, the gap between the physicians' healing intentions and the harm produced in patients and their relatives had been enormous (▶ Chap. 55, "Ethics in Psychiatry"). This general development was exacerbated largely through the research advances of the clinical neurosciences since the 1940s and 1950s, which have been examined in recent scholarly literature in the Science and Technology Studies (e.g., in Hayward 2001; Pickering 2004). The technological properties of the shock therapies and the neurostimulatory approaches make it clear that the invasive and manipulative operations on patients' central nervous system need to be regarded as two sides of the same biomedical coin: therapeutic technologies versus experimental instruments in tests with human subjects. It is precisely this amalgamation which adds further ethical questions to the problem of distinguishing between the epistemologies that clinical neuroscientists bring into their operation theaters, clinics, and test laboratories – for the better or worse (Roskies 2002).

By placing these developments in a comparative perspective, we see the often differing stances on patients' consent (Roelcke 2004), their relation to public perception (Reverby 2000), and the preparation of research funding requests to government agencies are all major factors in the development of new neuroethics concerns in the fields of psychiatry and clinical neuroscience (Dees 2004). In mapping out biomedical and technological advances, some significant connections, important co-developments, and continuing ethical problem areas become visible in this chapter. Furthermore, those ethical problems that relate to the changing historical relationship between the doctor and the patient raise central questions about the ethical status of informed consent in general. They point to the worrisome and often meandering course that reshaped the boundaries between neuroscientific research interests and clinical and therapeutic work (Farah and Wolpe 2004) (▶ Chap. 21, "What Is Normal? A Historical Survey and Neuroanthropological Perspective"). The individual historical examples from the rather recent and extended history of psychiatry and clinical neuroscience presented in this chapter make us aware of the continuingly problematic opposition between medical progress and patient ethics (Lederer 1995) as well as the intrinsically blurred conceptual boundary between clinical research, therapeutic intentions, and the autonomy of the individual patients within the framework of modern and ongoing biomedical progress (Heinrichs 2006) (▶ Chap. 25, "Impact of Brain Interventions on Personal Identity").

Cross-References

- ▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
- ▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
- ▶ Ethics in Neurosurgery
- ▶ Ethics in Psychiatry
- ▶ History of Psychopharmacology: From Functional Restitution to Functional Enhancement
- ▶ Human Brain Research and Ethics
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Informed Consent and the History of Modern Neurosurgery
- ▶ Neural Transplantation and Medical Ethics: A Contemporary History of Moral Concerns Regarding Cerebral Stem Cell and Gene Therapy
- ▶ Relationship of Benefits to Risks in Psychiatric Research Interventions
- ▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Ackerknecht, E. H. (1968). *A short history of psychiatry*. New York: Hafner.

Ackerknecht, E. H. (1986). Private institutions in the genesis of psychiatry. *Bulletin of the History of Medicine, 60*, 387–395.

Alesch, F., & Keith, M. (2004). Historische Einleitung. In J. Krauss & J. Volkmann (Eds.), *Tiefe Hirnstimulation* (pp. 1–10). Darmstadt: Dr. Dietrich Steinkopff.

Baumeister, A. (2000). The Tulane electrical brain stimulation programme: A historical case study in medical ethics. *Journal of the History of the Neurosciences, 9*, 262–278.

Bellak, L. (1948). *Dementia praecox: The past decade's work and present status: A review and evaluation*. New York: Grune & Stratton.

Borck, C. (2005). *Hirnstroeme. Eine Kulturgeschichte der Elektroenzephalographie*. Goettingen: Wallstein.

Brazier, M. A. B. (1988). *A history of neurophysiology in the 19th century*. New York: Raven.

Danzer, G. (1992). *Der wilde Analytiker. Georg Groddeck und die Entdeckung der Psychosomatik*. Munich: Fink.

Dees, R. H. (2004). Slippery slopes, wonder drugs, and cosmetic neurology. The neuroethics of enhancement. *Neurology, 63*, 951–952.

Delgado, J. (1969). *Physical control of the mind. Toward a psychocivilized society*. New York: Harper & Row.

Engstrom, E. (2004). *Clinical psychiatry in Imperial Germany. A history of psychiatric practice*. New York: Cornell University Press.

Engstrom, E. J., & Weber, M. (2007). Making Kraepelin history: A great institution? *History of Psychiatry, 18*, 267–273.

Farah, M. J., & Wolpe, P. R. (2004). New neuroscience technologies and their ethical implications. *The Hastings Center Report, 34*, 35–45.

Feindel, W. (1992). Brain physiology at the montreal neurological institute: Some historical highlights. *Journal of Clinical Neurophysiology, 9*, 176–194.

Feldman, R. P., Alterman, R. L., & Goodrich, J. T. (2001). Contemporary psychosurgery and a look to the future. *Journal of Neurosurgery, 95*, 944–956.

Finger, S. (2000). *Minds behind the brain: A history of the pioneers and their discoveries*. Oxford: Oxford University Press.

Fins, J. J. (2008). A leg to stand on: Sir William Osler and Wilder Penfield's 'neuroethics.' *The American Journal of Bioethics Neuroscience, 8*(1), 37–46.

Fradelos, C. K. (2008). The last desperate cure: Electrical brain stimulation and its controversial beginnings. PhD Dissertation, University of Chicago.

Gavrus, D. (2011). Men of dreams and men of action: Neurologists, neurosurgeons and the performance of professional identity, 1925–1950. *Bulletin of the History of Medicine, 85*, 57–92.

Glannon, W. (Ed.). (2007). *Defining right and wrong in brain science: Essential readings in neuroethics*. Washington, DC: The Dana Press.

Griesinger, W. (1845). *Pathologie und Therapie der psychischen Krankheiten*. Stuttgart: Krabbe.

Grob, G. N. (1983). *Mental illness and American Society. 1875–1940*. Princeton: Princeton University Press.

Hagner, M. (2000). *Homo cerebralis. Der Wandel vom Seelenorgan zum Gehirn*. Frankfurt am Main: Insel Press.

Harrington, A. (1996). *Reenchanted science, holism in German culture from Wilhelm II. to Hitler*. Princeton: Princeton University Press.

Hayward, R. (2001). The tortoise and the love machine: Grey Walter and the politics of electro-encephalography. *Science in Context, 14*, 615–641.

Heath, R. G. (1954). *Studies in schizophrenia*. Cambridge: Cambridge University Press.

Heinrichs, B. (2006). *Forschung am Menschen: Elemente einer ethischen Theorie biomedizinischer Humanexperimente*. Berlin: DeGruyter.

Karenberg, A. (2007). Klinische Neurologie in Deutschland bis zum Ersten Weltkrieg – die Begruender des Faches und der Fachgesellschaft. In D. Koempf (Ed.), *100 Jahre Deutsche Gesellschaft fuer Neurologie* (pp. 20–29). Berlin: De Gruyter.

Kraepelin, E. (1915). *Psychiatrie. Ein Lehrbuch fuer Studierende und Aerzte*. Leipzig: Barth.

Leahey, T. H. (2000). *A history of psychology. Main currents in pscyhological thought*. New Jersey: Prentice Hall.

Lederer, S. (1995). *Subjected to science: Human experimentation in America before the Second World War*. Baltimore: Johns Hopkins University Press.

Lindemann, M. (1999). *Medicine and society in early modern Europe*. Cambridge: Cambridge University Press.

Magoun, H. W. (2002). *American neuroscience in the twentieth century: Confluence of the neural, behavioral, and communicative streams* (Edited and annotated by Louise H. Marshall). Lisse: A. A. Balkema.

Morel, B. A. (1857–1858). *Traité des dégénérescence physique, et intellectuelles et morales de l'espèce humaine* (2 Vols.). Paris: Baillière.

Penfield, W. (1978). *The mystery of the mind*. Princeton: Princeton University Press.

Pickering, A. (2004). The science of the unknowable: Stafford Beer's cybernetic informatics. In M. E. Bowden & W. B. Rayward (Eds.), *The history and heritage of scientific information systems: Proceedings of the 2002 conference* (pp. 29–38). Medford: Information Today.

Porter, R. (1995). Shaping psychiatric knowledge: The role of the asylum. *Clio Medica, 29*, 255–273.

Porter, R. (2001). Nervousness, eighteenth and nineteenth century style: From luxury to labour. *Clio Medica, 63*, 31–49.

Pressman, J. (1998). *Last resort: Psychosurgery and the limits of medicine*. New York: Cambridge University Press.

Reiser, S. J. (1978). *Medicine and the reign of technology*. Cambridge: Cambridge University Press.

Reverby, S. (Ed.). (2000). *Tuskegee truths: Rethinking the tuskegee syphilis study*. Chapel Hill: University of North Carolina Press.

Richardson, D. E. (1980). Thalamic stimulation in the control of pain. *Southern Medical Journal, 73*, 283–285.

Roelcke, V. (2004). Historical perspectives on human subjects research during the 20th century, and some implications for present day issues in bioethics. In V. Roelcke & G. Maio (Eds.), *Twentieth century ethics of human subjects research: Historical perspectives on values, practices and regulations* (pp. 11–18). Stuttgart: Franz Steiner.

Roelcke, V. (2005). Continuities or ruptures? Concepts, institutions and context of twentieth-century German psychiatry and mental health care. In M. Gijswijt-Hofstra, H. Oosterhuis, J. Vijselaar, & H. Freeman (Eds.), *Psychiatric cultures compared. Psychiatry and mental health care in the twentieth century* (pp. 162–182). Amsterdam: Amsterdam University Press.

Roskies, A. (2002). Neuroethics for a new millennium. *Neuron, 35*, 21–23.

Rousseau, G. (1990). *The languages of psyche: Mind and body in enlightenment thought*. Berkeley: California University Press.

Sammet, K. (1997). *Ueber Irrenanstalten und deren Weiterentwicklung in Deutschland: Wilhelm Griesinger im Streit mit der konservativen Anstaltspsychiatrie 1865–1868*. Hamburg: LIT Press.

Schmiedebach, H. P. (2001). The public's view of neurasthenia in Germany – Looking for a new rhythm of life. In R. Porter & M. Gijswijt (Eds.), *Cultures of neurasthenia. From Beard to the First World War* (pp. 219–238). Amsterdam: Rodopi.

Shorter, E. (2009). The history of lithium therapy. *Bipolar Disorders, 2* (Suppl), 4–9.

Slattery, J. P. (1990–1991). *Report of the royal commission into deep sleep therapy*. Sydney: New South Wales Parliamentary Paper 304.

Stahnisch, F. W. (2007). Griesinger, Wilhelm (1817–1869). In W. F. Bynum & H. Bynum (Eds.), *Dictionary of medical biography* (pp. 528–583). Westport: Greenwood.

Stahnisch, F. W. (2012). *Medicine, life and function: Experimental strategies and medical modernity at the intersection of pathology and physiology*. Bochum: Projektverlag.

Stone, J. L. (2001). Dr. Gottlieb Burckhardt – The pioneer of psychosurgery. *Journal of the History of the Neurosciences, 10*, 79–92.

Temkin, O. (1955). Therapeutic trends and the treatment of syphilis before 1900. *Bulletin of the History of Medicine, 29*, 309–316.

Tone, A. (2005). Listening to the past: History, psychiatry, and anxiety. *Canadian Journal of Psychiatry, 50*, 373–380.

Valenstein, E. S. (1986). *Great and desperate cures: The rise and decline of psychosurgery and other radical treatments for mental illness*. New York: Basic Books.

Vidal, F., & Ortega, F. (2011). Are there neural correlates of depression? In S. Choudhury & J. Slaby (Eds.), *Critical neuroscience: A handbook of the social and cultural contexts of neuroscience* (pp. 345–366). Oxford: Wiley-Blackwell.

von Weizsaecker, V. (1930). *Soziale Krankheit und soziale Gesundheit*. Berlin: Springer.

Weisz, G. (1994). The development of medical specialization in nineteenth–century Paris. In A. La Berge & M. Feingold (Eds.), *French medical culture in the nineteenth century* (pp. 149–187). Amsterdam: Rodopi.

Historical and Ethical Perspectives of Modern Neuroimaging

33

Fernando Vidal

Contents

Abstract

Both in its development and in the definition of its tasks, neuroethics has been intimately connected to neuroimaging, especially to the widespread application of functional brain imaging technologies such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). Neuroimaging itself, in particular its uses, interpretation, communication, media presence, and public understanding, has been one of neuroethics' primordial subjects. Moreover, key neuroethical issues, such as brain privacy or the conceptualization of blame, responsibility, and in general human personhood, have largely gained from neuroimaging the form under which neuroethics deals with them. The use of neuroimaging techniques to investigate phenomena usually associated with research in the humanities and human sciences brought those phenomena into the orbit of neurobiological explanation. Neuroethics emerged in the context of such technology-driven intellectual and professional developments. Thus, more

F. Vidal
ICREA (Catalan Institution for Research and Advanced Studies), CEHIC (Center for the History of Science, Autonoma University of Barcelona), Barcelona, Spain

Unitat d'Història de la Medicina, Facultat de Medicina, M6/130, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain
e-mail: fernando.vidal@icrea.cat

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_27,

than an important stimulus for the development of neuroethics or a particular source of neuroethical challenges, the spread of functional neuroimaging can be considered as a condition of possibility of the field. In return, neuroethics supports the claim that the neurosciences, in particular by way of functional neuroimaging, will revolutionize "traditional" ways of understanding the human. To the extent that such a claim is debatable, neuroethics might benefit from examining its special relationship to neuroimaging.

Introduction

Unheard of in the 1990s, neuroethics required just a few years in the early 2000s to gain autonomy from bioethics, professionalize itself, and create networks, platforms, societies, journals, institutes, teaching, and research programs (Hoyer 2010; Conrad and De Vries 2012). Its main explicit goal is to examine and anticipate the ethical, social, and legal consequences of neuroscientific knowledge and its applications as well as to contribute to the ethics of neuroscience. In its development, history, and self-definition, neuroethics has been inherently connected to neuroimaging, especially to the increasingly widespread employment of functional brain imaging technologies such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI).

On the one hand, neuroimaging itself, as well as its interpretation, communication, media presence, and public understanding, has been one of neuroethics' primordial subjects. On the other hand, key neuroethical issues, such as "brain privacy," or the conceptualization of blame, responsibility, and in general human personhood, have gained from neuroimaging the form under which neuroethics deals with them. The use of neuroimaging techniques to investigate phenomena usually associated with research in the humanities and human sciences brought those phenomena into the orbit of neurobiological explanation. Neuroethics emerged in the context of such technology-driven intellectual developments. Thus, more than an important stimulus for the development of neuroethics or a particular source of conceptual and practical challenges, the spread of functional neuroimaging can be considered as a condition of possibility of the field.

Neuroimaging and Neuroethics: A Special Relationship

Coinciding with the "Decade of the Brain," the number of fMRI studies published in peer reviewed journals exploded from 1991 to 2001, and so did the number of journals publishing neuroimaging research. In parallel, from a small number of articles, mostly on sensory and motor tasks, the application of fMRI to human-science topics grew to 865 in 2001, with an average increase of 61 % per year during that period (Illes et al. 2003). By the end of 2004, 3,824 more articles had been published (Racine et al. 2005), and their coverage in the printed media

displayed the same trend (Illes et al. 2006) and has remained broadly optimistic (Racine et al. 2010b). By the middle of the decade, there was an avalanche of neuroimaging studies on topics with potential ethical, legal, social, and policy implications in four main areas: (1) altruism, empathy, decision-making, cooperation, and competition; (2) judging faces and races; (3) lying and deception; and (4) meditating and religious experience (Illes et al. 2006). Commercial products using neuroimaging results or devices appeared in the same period: major companies pay for neuromarketing services (Burkitt 2009); neurobics entrepreneurs offer resources for brain maintenance and enhancement purposes (Ortega 2011); private firms claim to have developed operative fMRI lie detectors (see Bizzi et al. 2009). Such scientific and extrascientific developments are a major motivation and source material for neuroethics (Racine et al. 2010a).

The expansion of neuroimaging and its connection to neuroethics can also be charted through the emergence of institutional and investigative programs whose names, from neuroaesthetics to neurotheology through neuroanthropology, neuroeconomics, neuroeducation, neurohistory, neurolaw, neuropsychoanalysis, and others, combine the prefix *neuro* with the name of a human or social science. Within these sciences, the "neuroscientific turn" (Feuerhahn and Mandressi 2011; Littlefield and Johnson 2012) occupies only a minority of professionals. Yet it contributes to motivate neuroethics. The new *neuro* disciplines, whose main tool is fMRI, aim to elucidate neurobiologically the phenomena they study. Moreover, they tend to assume that humans are essentially their brains, that ultimate explanations about human individual and collective behavior must be provided by brain research, and that such explanations challenge "traditional" beliefs about human nature and radically transform our views about it.

On the one hand, neuroethics is concerned with the results of those new research areas. On the other hand, as a field, it shares their assumptions, which it invoked in its successful effort to be considered a discipline separate from bioethics (see *American Journal of Bioethics*, 5(2), 2005). Its reasons for asserting autonomy were the same as those that underlie the "neuroscientific turn," namely, the existence of an "intimate connection" between brain and behavior, the "peculiar relationship between our brains and ourselves," and "the intuition that our ever-increasing understanding of the brain mechanisms underlying diverse behaviors has unique and potentially dramatic implications for our perspective on ethics and social justice" (Roskies 2002, p. 21). Neuroethicists affirm that the neurosciences will redefine "our sense of selfhood and brain-body relations" (Wolpe 2002, p. 8), metamorphose ancient philosophical questions about personhood, and give rise to new moral and legal challenges (Illes and Racine 2005). Given "the capacity of imaging technology to animate mind-brain relationships in ways that compel us to reexamine concepts of identity, personal responsibility, and criminal culpability," neuroimaging turns out to be "a logical focus" of neuroethical inquiry (Kulynych 2002, p. 345). Beyond the connection of brain and personal identity, the case for "neuro exceptionalism" is bolstered by futuristic claims about the unique problems that, according to neuroethicists, functional neuroimaging will raise (see counterarguments in Schick 2005).

The special relationship between neuroethics and functional neuroimaging derives precisely from the latter's capacity to measure brain activity during the performance of an experimental task. The title of an early programmatic article highlights the extent to which neuroethics exists "in a new era of neuroimaging" (Illes 2003); that of another, "from neuroimaging to neuroethics," underlines the origins of the new discipline in the applications of functional imaging (Illes et al. 2003). A 2010 article on possible approaches to revising the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) asked, "What happens if dramatic technical breakthroughs in genetics, imaging or neuroscience cast the problems of psychiatric nosology in an entirely new light?" (Kendler and First 2010). In spite of many caveats and open questions about the nosological and clinical relevance of neuroimaging, the question epitomizes one of the expectations that justify neuroethics: neuroimaging *will* have a significant impact on diagnosis and treatment (Malhi and Lagopoulos 2008; Farah and Gillihan 2012).

In addition to hundreds of articles, the seven neuroethics multiauthored volumes (one handbook, three anthologies, and two collections of new essays) published between 2006 and 2012 confirm the special position of functional neuroimaging in the neuroethical landscape (Illes 2006; Glannon 2007; Giordano and Gordijn 2010; Farah 2010a; Illes and Sahakian 2011; Chatterjee and Farah 2013). In sum, to a very large extent, the existence of neuroethics as a research program with normative aspirations derives from and depends on the expanding application range of functional neuroimaging, coupled with the assumption that "imaging the brain provides information about the mind" (Farah 2010b, p. 4).

Illustrating, Mapping, and Imaging

As far as fMRI is concerned, the assertion that "imaging the brain provides information about the mind" implies that correlations between cerebral activity and the realization of an experimental task during scanning indeed capture a significant relationship between physical brain states and subjective mental states. During the control and experimental conditions, fMRI measures BOLD ("blood-oxygen level-dependent") signals (a hemodynamic proxy of neuronal activity), which are then subtracted from each other; the difference defines the correlation between neural activity and the function which the experimental task is supposed to involve. This neural correlation, usually presented as the "activation" of a brain area, embodies the "information" imaging provides about the mind.

The subtraction method and what that "information" amounts to have long been in dispute (Van Orden and Paap 1997; Kosslyn 1999; Hardcastle and Stewart 2002; Coltheart 2006; Roskies 2009, 2010). It is nonetheless common to say that the "activated" areas "underlie" or "underpin" the functions being studied (Schleim and Roiser 2009). Thus, even though fMRI results are correlations, they are generally discussed in terms that strongly suggest causality (Vidal and Ortega 2011), and this is also part of the "information" imaging is believed to furnish. The colored portions in scans, which are only graphic renderings of correlational results, are widely

perceived as displaying the brain areas that become active during the experimental tasks and sustain the function being investigated.

The methods whereby fMRI images are produced, the information they provide, and the expectations they generate dictate their place in the history of representations of the brain (Clarke and Dewhurst [1972] 1996; Larink 2011). These representations, which have been integral to the production of knowledge about the brain (Mandressi 2011), can be divided into three classes: images that realistically depict morphology, images that show localization of function in anatomical structure, and images that result from records of function.

Descriptive anatomy, which is mostly based on dissection, is accompanied by images of the first class, which may be called "illustrations." Examples are innumerable, but since this class is historically the earliest, it is fitting to mention an early case: Christopher Wren's superb etchings of Thomas Willis' *Cerebri anatome* (1664). Wren's images look realistic; however, they are not like photographs of a dissected brain, but synthesize observations made on different brains (Flis 2012). Their accuracy and the quality of the knowledge they convey contribute to explain why they are considered epoch-making. However, when Willis (1621–1675), a major figure in the history of neuroanatomy (Molnár 2004; Zimmer 2004), dealt with the physiology of mind, he adhered to Galenic humoralism, which, unrelated to neuroscientific research, dominated medical thought until the eighteenth century (Temkin 1973).

According to Galenism, as the blood passes through the organs of the body, it is transformed into subtile fluids or "spirits." It first becomes a "natural spirit" responsible for nutrition and growth. In the heart, the natural spirit mixes with air from the lungs and turns into the "vital spirit" on which motor and vital functions depend. A final transformation takes place in the cerebral ventricles with the formation of the "animal spirits" necessary for sensitive and intellectual functions. The animal spirits were believed to reside in and move among the brain ventricles, which thus constituted the seat of mental faculties. From front to back of the head, these were the "common sense," the imagination and fantasy, the judgment and intellect, and memory (Harvey 1975; Kemp 1990). This "cell doctrine" was the object of numerous illustrations in the medieval and early modern periods. Willis studied the solid parts of the brain. At the same time, however, he linked the imagination to an "undulation" of the spirits from the center of the brain towards its circumference and speculatively placed its seat in the corpus callosum; he made memory depend on the movement of the spirits from the periphery towards the center of the brain and located it in the cortex; and he situated sensory coordination in the corpus striatum, which received the impressions going towards the brain and was, he conjectured, the path by which the animal spirits descended towards the extremities.

A second class of brain images shows functions as localized in anatomical structure; they are constructed by mapping anatomo-clinical or experimental results onto an anatomical picture. Such morphological "mappings," as they may be called, began developing in the eighteenth century. Brain research between the late seventeenth and early nineteenth century was closely connected to investigations of the structure and function of the sense organs. Since the nerves united the soul

and the body, as well as the external world and the brain, the nervous system became in the Enlightenment the common ontological matrix of the sciences of the body and the sciences of the soul (Figlio 1975; Rousseau 2007; Vidal 2011). Throughout the eighteenth century, anatomical knowledge of the brain advanced through surgical practice and pathological and microscopic studies following vivisection or unintended lesions, new dissection, fixation and staining methods, and, starting in mid-century, electrical stimulation experiments (Barbara 2008).

The early decades of the nineteenth century saw the spread of phrenology (Clarke and Jacyna 1987; Renneville 2000; van Wyhe 2002). Based on the theories of the Viennese physician Franz Joseph Gall (1758–1828), who called it "organology" and "doctrine of the skull," it assumed that the brain is the organ of the mind, that each faculty has its own brain "organ," that the size of each organ is proportional to the strength of the corresponding faculty, that the brain is shaped by their differential growth, and finally, that since the skull owes its form to the underlying brain, its "bumps" reveal individual aptitudes and tendencies. Phrenology, and the accompanying practices of cranioscopy and cranial palpation, remained hugely popular into the 1840s, and phrenological publications appeared steadily until after World War I. The typical picture shows a head with neatly delimited zones marked on it, often resulting in a grid, each corresponding to a faculty of the mind.

In the same period, experimental psychophysiology and pathological anatomy gave impulse to the project of localizing mind in the brain (Star 1989; Finger 1994). While phrenology correlated dispositions with cranial shape, the anatomo-clinical method searched for correspondences between symptoms and brain lesions. The advocates of situating mental faculties in discrete loci and those who insisted on the unity of intelligence and the integrated nature of brain action shared such methodological orientation. The case of "Tan," an aphasic patient studied by the French anatomist and physical anthropologist Paul Broca (1824–1880), is paradigmatic of mid-nineteenth-century localization debates. Tan's clinical history and the postmortem study of his brain led Broca to conclude that the faculty of articulate language was possibly located in the second or third frontal convolution. In his foundational "Remarques sur le siége de la faculté du langage articulé," Broca (1861) identified correspondences between mental functions and brain areas. He localized the higher "brain faculties," including judgment, reflection, and abstraction, in the frontal lobes, and the feelings and passions in the temporal, parietal, and occipital lobes.

For nineteenth-century British and German brain scientists, the method of correlating clinical and pathological phenomena was suspiciously reminiscent of the craniological approach (Young 1990). Few, however, would have denied that the extraordinary positive or negative qualities of geniuses, criminals, and the mentally ill were somehow inscribed in the brain. This brand of localizationism matched the nineteenth-century development of anthropometry and the related elaboration of physiognomic, cranial, and bodily typologies (Hagner 2004; Rafter 2008).

Beyond national differences, to know the brain was to know what its parts did and were responsible for. By the late nineteenth century, cerebral localization, differentiation of function, and the correlation of structure and function had become investigative principles; the physiological and functional study of discrete regions of the cortex (mainly by means of lesions and electrical stimulation) resulted in detailed anatomical cortical maps. In the twentieth century, clinical and experimental methods came increasingly together. For instance, as a neurosurgeon treating epileptics with open-skull surgery, Wilder Penfield (1891–1976) stimulated sections of the cortex. On the basis of patients' responses, he mapped areas corresponding to motor and sensory functions and represented them as a "homunculus" whose features are drawn proportionally to the associated brain areas (Penfield and Boldrey 1937; Penfield and Rasmussen 1950; Pogliano 2012). Although the homunculus does not look like a brain and is not topographically accurate, it belongs to the same class of images as those drawn by the neurologist and neurosurgeon Otfrid Foerster (1873–1941), with whom Penfield studied and collaborated, which map sensorimotor regions discovered by means of accidental or experimental lesions onto a schematic but realistic depiction of the cortex.

Illustrations and mappings differ from what might be called "imagings," which consist of tracings derived from recording in vivo processes. The best known is electroencephalography, but starting in the 1860s, the earliest methods attempted to measure and graphically record regional head or brain temperatures, as well as the amount of blood circulating in the brain, and to correlate them with cognitive tasks; research in the 1890s established relations between cerebral functional activity, energy metabolism, and blood flow (Zago et al. 2012). The term *imaging* suits these different visualization techniques, since the progressive aspect indicated by the *-ing* form denotes an ongoing action.

In the early twentieth century, self-recording instruments for graphically registering physiological activity gave rise to the hope of establishing psychological processes as physical facts and of generating a "brainscript" from which to read psychic phenomena directly (Borck 2005). In the 1930s, although electroencephalography inventor Hans Berger (1873–1941) conceptualized EEG "as a device for documenting psychic processes via their physiological correlates" (Borck 2001, p. 583), the technique was celebrated as a mind-reading device. Later digitization has not fundamentally altered the hopes and the issues connected to imaging. Instead, like earlier technologies, functional neuroimaging has been widely associated with mind reading.

The form of localization which functional imaging produces does not consist of establishing the "seat" of particular functions but of showing areas and circuits somehow involved in those functions. Neuroimaging thus collapses the distinction between morphological and functional images and acts as a sort of inward physiognomy (Hagner 2009). Essential for neuroethical considerations is the fact that, even when they look like the realistic reproductions of an entire brain or of traditional brain sections, fMRI images neither apply optical principles nor reproduce anything: they are "belief-opaque in a way photography is not" (Roskies 2007, p. 871). Their appearance embodies decisions about how to generate, process, and

represent numerical data recorded from a functioning brain; this data could be given the form of graphs or curves rather than vibrantly colored brains.

Insofar as fMRI brings to light functional integration, it replaces the static localizationism of mappings with the dynamic localizationism of imaging. From the historical and epistemological viewpoints, it is nevertheless significant that the basic beliefs and hopes associated with functional neuroimaging as well as many of the uses to which it has been put have led to comparisons with phrenology (Uttal 2001) and that phrenology and neuroimaging have been said to share misguided structure-function mapping strategies (Poldrack 2010).

The parallels between phrenology and neuroimaging also evince neuroethical concerns (Tovino 2007b, part V). Both moved rapidly from clinical and research contexts to the public arena, especially law and education, as well as to commercial applications in personal decision-making and employment screening. The issues they raise are related to localization, but not specifically to *functional* neuroimaging. While phrenology professed to identify the cerebral seat of particular functions and anatomo-pathology actually enabled their localization, functional imaging generates correlations between cerebral activity and the performance of a mental task. Yet, although neuroethicists and neuroscientists tend to believe that thanks to fMRI the mind/brain sciences "have made more progress in a decade than had been made in over two thousand years" (Cowey 2001, p. 254), the new functional visualization techniques coexist with old psychological parameters (Hagner and Borck 2001, p. 508) as well as with traditional formulations of the moral and philosophical problems which those techniques allegedly revolutionize.

The Structural and the Functional

As seen above, neuroimaging is part and parcel of the "neuro exceptionalism" that legitimizes the autonomy of neuroethics. This does not mean that all major neuroethical topics are inherently connected to neuroimaging. In the case of the pharmacological enhancement of normal neurocognitive functions, fMRI is sometimes used, but effects are typically assessed with controlled clinical trials comparing the experimental drug with placebo (Repantis et al. 2010). The related neuroethical discussion focuses on classical bioethical issues, such as safety, freedom, fairness, and regulation (Farah et al. 2004; Glannon 2008; Illes and Sahakian 2011; Chatterjee and Farah 2013).

The same applies to some of the central neuroethical topics that are directly connected to brain imaging. First example is incidental findings (Detre and Bockow 2013). Such findings are common in human subject research. In neuroimaging studies, where they display a relatively high prevalence of 2–3 % in neurologically asymptomatic people, they may reveal potentially symptomatic or treatable conditions as well as various markers of cerebrovascular disease (Morris et al. 2009). Reporting them may lead to life-saving care; however, since neuroimaging research often involves scans which are not clinical grade and which may be read by researchers with no training in diagnostic interpretation, suspicious findings may

turn out to be false positives. Such a situation involves dilemmas qualitatively identical to those that arise in bioethics. What makes incidental findings in neuroimaging research a particularly significant topic for neuroethics is that brain anomalies may threaten "the continued function of the individual" (Wolf 2011, p. 626).

Second example is privacy rights. Here again, neuroimaging occupies a prominent place. Its implications offer analogies to those of HIV test results and genetic information, which in many legal systems are subject to heightened confidentiality protection measures. Here too the *neuro* lacks specificity. Moreover, there would be no reason to apply extraordinary measures to data which, though obtained by way of imaging, chiefly corroborates otherwise diagnosed conditions or even provides "correlates" of such psychiatric disorders as depression or schizophrenia. Neuro-exceptional measures could be defended – but that would require accepting that fMRI has the "potential to reveal insights about an individual's thoughts, feelings, preferences, prejudices, and other social characteristics and behaviors" (Tovino 2007a, p. 487).

Third example is neuroscience and the law (Freeman 2011; Spranger 2012). The issues examined in this context, sometimes under the appellation "neurolaw," include among others neuroimaging lie detection, the admissibility of brain scans in the courts, or their relevance for determining degrees of guilt and assessing responsibility. As in neuroethics at large, discussions are very largely focused on or derived from the forensic uses of brain imaging (Simpson 2012a). They typically combine calls for caution concerning real-world applications of neuroimaging with the conviction that, as it becomes "more reliable, standardized, and informative, attempts to use its results in civil and criminal proceedings of all types will increase dramatically" (Simpson 2012b, pp. xv–xvi). In fact, there is no agreement on whether or not neuroimaging will affect legal arguments about free will or undermine agency and responsibility (Aggarwal 2009).

As stated above, neuroimaging collapses the distinction between morphological and functional images. Yet, as the examples of incidental findings, privacy rights, and neuroscience and the law demonstrate, neuroethics relates most specifically to perceptions about the impact of its *functional* capabilities. These, however, have fallen short of their promise. For example, since the introduction of computerized tomography (CT) in the early 1970s, structural neuroimaging has helped to identify pathological lesions in epileptic patients. In contrast, functional imaging has yielded results of very limited clinical significance, "despite many research efforts and massive funding" as well as claims about its potential to localize function or aid in presurgical localization (Shorvon 2009, pp. 47–48).

In the domain of schizophrenia, despite comparable investments, since the first computerized axial tomographies were performed in the mid-1970s, "no consistent or reliable anatomical or functional alterations have been unequivocally associated with psychosis or schizophrenia and no clinical applications have been developed" (Borgwardt and Fusar-Poli 2012, p. 270). The hopes, however, remain high that combining imaging modalities among themselves and with neurodevelopmental

and genetic data will yield unequivocal anatomical and physiological markers (e.g., Insel 2010; Calhoun and Hugdahl 2012).

In the domain of mood disorders, the most consistent structural neuroimaging finding is white matter hyperintensities in unipolar patients; relatively consistent functional neuroimaging findings include a decreased anterior paralimbic and cortical activity in depressed or bipolar patients. Studies in the area repeatedly affirm that brain imaging will contribute to understanding pathogenesis (Koolschijn et al. 2009). A meta-analytic study, however, concluded that after 25 years of scanning bipolar disorder patients and generating over 7,000 MRIs, the consistent findings are few and "regionally nonspecific" (Kempton et al. 2008).

In spite of the extremely limited substance of the empirical results, their lack of clinical significance, the interpretive difficulties they present, and the many methodological and epistemological problems they raise (Vidal and Ortega 2011), researchers involved in the neuroimaging of psychiatric disorders conclude that "fMRI has emerged as an exciting and pivotal tool that has enabled investigators to probe the mind and anchor their findings within the brain" (Malhi and Lagopoulos 2008, p. 114). Even the most cautious discussions (Farah and Gillihan 2012) consider it merely "premature" to include brain imaging among diagnostic criteria for the most recent Diagnostic and Statistical Manual of Mental Disorders (DSM V). This contrasts with other opinions, which recognize that, on account of the very nature of cerebral functions, imaging "may be an inherently inappropriate technology" to study them (Shorvon 2009, p. 48).

The Practical and the Philosophical: Future Directions

As neuroethicist Martha Farah (2005) pointed out, some neuroethical questions concern the practical implications of neurotechnologies for individuals and society, while others, more philosophical, concern the relationship of mind and brain as well as the material determinants of human behavior. The "practical" are mainly related to neuroenhancement, which raises questions about fairness and equal opportunity, as well as to the uses of neuroimaging, which raise issues of privacy, responsibility, and public policy in domains such as the law, education, and public health. The "philosophical," derived from more basic neuroscience but also largely dependent on neuroimaging, touch on the ways we think about ourselves as persons and moral agents.

Overall, the way neuroethics has been practiced up to the moment this chapter was written corroborates the observation that

> neuro-ethical anxieties [about for instance enhancement, authenticity or free will] have become part of the very problem they seek to address. In inflating the powers of the new brain sciences, [. . .] they have become part of a culture of hype and hope, futurology and fear. Far from opening a space for clear thinking about the issues that confront us today, much neuro-ethics implicitly makes the case for those who live on these expectations, be they researchers in search of grants, corporations in search of investment or popular science writers who thrive on sensationalism to sell their products [. . .]. (Singh and Rose 2006, p. 100)

Similarly, mainstream neuroethical practice confirms that "[a]s much as neuroethicists claim to guard society from neuroscience's abuses, they also guard neuroscience from society's criticisms. This enables neuroscientists to more freely push the boundaries on new technologies, piquing public curiosity and expanding the boundaries of neuroethics" (Conrad and De Vries 2012, p. 320). In no area is this truer than in neuroethicists' handling of neuroimaging.

Hence the following two possible new directions for neuroethics. One would consist of self-reflexively examining its role as expert consultant, linked to its focus on "the practical." To extrapolate from criticisms addressed to bioethics (John 2004, pp. 178, 179), it may be observed that such focus concerns problems that are dictated mainly by North American markets and may be "of little ethical concern in the larger context." The predominant empiricist approach, which seeks to assess dilemmas raised in the field, "distorts ethical inquiry by making us focus on particular cases which may not be those of most ethical importance." While it is meaningful to discuss the impact of imaging technologies on brain privacy, it might be more significant to ask whether it is socially responsible to allocate so many resources to neuroimaging in areas where the technology might be unsuitable for the goals pursued, and where most results have long been, and may be expected to remain of doubtful value.

A new task for neuroethics would consist of examining the beliefs and interests underlying such a situation, in order to explain and eventually question its opportunistic alliance with the neuroimaging-driven turn in psychiatry and the human sciences. For example, in dealing with "the puzzle of neuroimaging and psychiatric diagnosis" (Farah and Gillihan 2012) – namely, that although image-based brain correlates have been documented for most psychiatric disorders, functional imaging plays almost no role in nosology, etiology, and diagnosis – neuroethicists warn against premature applications and inflated claims but expect that future advances will allow imaging to play a key role in those areas. An alternative would be to place their "expectational discourses" (Brosnan 2011) and the *neuro* pursuits they legitimate, in socioeconomic, political, historical, and ethical perspectives (Cooter 2010a, b).

Consequently, a second possible new direction for neuroethics would include rethinking its approach to "the philosophical," which has so far embraced an exalted assessment of the potentialities of neuroimaging. The editorial to the first issue of the journal *Neuroethics* asserted that "[n]euroscientific knowledge promises to transform our understanding of [. . .] what it means to be a thinking being" and explained that, whereas medical advances "deal with our bodies, neuroscientific discoveries promise – or threaten – to reveal the structure and functioning of our minds, and, therefore, of our souls" (Levy 2008, p. 2). The "knowledge" and the "discoveries" alluded to are largely attributed to neuroimaging. Since such claims are the cornerstone of neuroethics, the discipline has taken them for granted, rather than dealing with them in their proper contexts.

Historically, however, the notion that, as persons, humans are essentially their brains, dates at the earliest from the late seventeenth century and goes hand in hand with "possessive individualism" and an understanding of personhood that emphasizes individual autonomy; contrary to what neuroethicists seem to believe, it does not result from neuroscientific advances but predates them and has contributed to

motivate brain research (Vidal 2009). History, which is absent from neuroethical inquiry except in the most cursory manner, suggests that trying to understand personhood by showing that there is an autonomous person network in the brain (Farah and Heberlein 2007) is both scientifically and philosophically unpromising. Similarly, sociological research into individuals' uses of "neuro" discourses (O'Connor and Joffe 2013) considerably undermines neuroethics' understanding of neuroscience's social "impact." Finally, a rich literature about the emergence of the brain as icon of Western modernity mediated by imaging technologies (e.g., Dumit 2004; Joyce 2008) problematizes neuroethicists' belief that "imaging the brain provides information about the mind." The present and future of neuroethics are tied to those of that icon in ways neuroethics itself would benefit from examining.

Cross-References

▶ Determinism and Its Relevance to the Free-Will Question
▶ Feminist Neuroethics: Introduction
▶ Human Brain Research and Ethics
▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
▶ Neuroethics and Identity
▶ Neuroimaging Neuroethics: Introduction
▶ Neurolaw: Introduction
▶ Neuromarketing: What Is It and Is It a Threat to Privacy?
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play
▶ Neuroscience, Neuroethics, and the Media

References

Aggarwal, N. K. (2009). Neuroimaging, culture, and forensic psychiatry. *Journal of the American Academy of Psychitry and Law, 37*, 239–244.

Barbara, J. G. (2008). Diversité et évolution des pratiques chirurgicales, anatomiques et physiologiques du cerveau au XVIIIe siècle. In C. Chérici & J.-C. Dupont (Eds.), *Les querelles du cerveau. Comment furent inventées les neurosciences* (pp. 19-54). Paris: Vuibert.

Bizzi, E., et al. (2009). *Using imaging to identify deceit: Scientific and ethical questions*. Cambridge, MA: American Academy of Arts and Sciences.

Borck, C. (2001). Electricity as a medium of psychic life: Electrotechnological adventures into psychodiagnosis in Weimar Germany. *Science in Context, 14*(4), 565–590.

Borck, C. (2005). *Hirnströme: Eine Kulturgeschichte der Elektroenzephalographie*. Berlin: Wallstein.

Borgwardt, S., & Fusar-Poli, P. (2012). Third-generation neuroimaging in early schizophrenia: Translating research evidence into clinical utility. *The British Journal of Psychiatry, 200*, 270–272.

Broca, P. (1861). Remarks on the seat of the Faculty of Articulated Language, following an observation of aphemia (loss of speech) (trans: Green, C.). http://psychclassics.yorku.ca/Broca/aphemie-e.htm. First published in *Bulletin de la Société Anatomique*, *6*, 330–357.

Brosnan, C. (2011). The sociology of neuroethics: Expectational discourses and the rise of a new discipline. *Sociology Compass, 5/4*, 287–297.

Burkitt, L. (2009). Neuromarketing: Companies use neuroscience for consumer insights. *Forbes*, 16 November. http://www.forbes.com/forbes/2009/1116/marketing-hyundai-neurofocus-brain-waves-battle-for-the-brain.html. Last Accessed 8 Dec 2012.

Calhoun, V. D., & Hugdahl, K. (2012). Cognition and neuroimaging in schizophrenia. *Frontiers in Human Neuroscience, 6*, Article 276.

Chatterjee, A., & Farah, M. J. (Eds.). (2013). *Neuroethics in practice: Medicine, mind, and society*. New York: Oxford University Press.

Clarke, E., & Dewhurst, K. E. (1996 [1972]). *An illustrated history of brain function: Imaging the brain from antiquity to the present* (2nd edn, with a new chapter by M. J. Aminoff). San Francisco: Norman.

Clarke, E. & Jacyna, L. S. (1987). *Nineteenth-century origins of neuroscientific concepts*. Berkeley: University of California Press.

Coltheart, M. (2006). Perhaps functional neuroimaging has not told us anything about the mind (so far). *Cortex, 42*, 422–427.

Conrad, E. C., & De Vries, R. (2012). Field of dreams: A social history of neuroethics. In M. Pickersgill & I. Van Keulen (Eds.), *Sociological reflections on the neurosciences* (pp. 299–324). Bingley: Emerald Group Publishing.

Cooter, R. (2010a). Inside the whale: Bioethics in history and discourse. *Social History of Medicine, 23*, 662–672.

Cooter, R. (2010b). *Neuroethical brains, historical minds, and epistemic virtues*. Unpublished lecture, Wellcome Trust Center for the history of medicine symposium “neuroscience and human nature.” http://www.ucl.ac.uk/histmed/downloads/articles/neuroethical_brains_historical_minds_epistemic_virtues.pdf

Cowey, A. (2001). Functional localization in the brain: From ancient to modern. *Psychologist, 14*, 250–254.

Detre, J., & Bockow, T. B. (2013). Incidental findings in neuroimaging studies. (In Chatterjee and Farah 2013, pp. 120–127).

Dumit, J. (2004). *Picturing personhood: Brain scans and biomedical identity*. Princeton: Princeton University Press.

Farah, M. J. (2005). Neuroethics: The practical and the philosophical. *Trends in Cognitive Neuroscience, 9*, 34–40.

Farah, M. J. (Ed.). (2010a). *Neuroethics: An introduction with readings*. Cambridge, MA: MIT Press.

Farah, M. J. (2010b). Neuroethics: An overview. (In Farah 2010a, pp. 1–9).

Farah, M. J., & Gillihan, S. J. (2012). The puzzle of neuroimaging and psychiatric diagnosis: Technology and nosology in an evolving discipline. *AJOB Neuroscience, 3*(4), 1–11.

Farah, M. J., & Heberlein, A. S. (2007). Personhood and neuroscience: Naturalizing or nihilating? *The American Journal of Bioethics, 7*(1), 37–48.

Farah, M. J., et al. (2004). Neurocognitive enhancement: What can we do and what should we do? *Nature Reviews Neuroscience, 5*, 421–425.

Feuerhahn, W., & Mandressi, R. (Eds.). (2011). *Les sciences de l’homme à l’âge du neurone* (special issue of *Revue d’histoire des sciences humaines*, No. 25).

Figlio, K. M. (1975). Theories of perception and the physiology of mind in the late eighteenth century. *History of Science, 12*, 177–212.

Finger, S. (1994). *Origins of neuroscience: A history of explorations into brain function*. New York: Oxford University Press.

Flis, N. (2012). Drawing, etching, and experiment in Christopher Wren’s figure of the brain. *Interdisciplinary Science Reviews, 37*(2), 145–160.

Freeman, M. (Ed.). (2011). *Law and neuroscience*. New York: Oxford University Press.

Giordano, J., & Gordijn, B. (Eds.). (2010). *Scientific and ethical philosophical perspectives in neuroethics*. New York: Oxford University Press.

Glannon, W. (Ed.). (2007). *Defining right and wrong in brain science: Essential readings in neuroethics*. Washington, DC: The Dana Press.

Glannon, W. (2008). Psychopharmacological enhancement. *Neuroethics, 1*, 45–54.

Hagner, M. (2004). *Geniale Gehirne. Zur Geschichte der Elitenhirnforschung*. Berlin:Wallstein.
Hagner, M. (2009 [2006]). The mind at work: The visual representation of cerebral processes (Trad. U. Froese.). In R. van de Vall & R. Zwijnenberg (Eds.), *Body within: Art, medicine and visualization* (pp. 67–90). Leiden: Brill.
Hagner, M., & Borck, C. (2001). Mindful practices: On the neurosciences in the twentieth century. *Science in Context, 14*(4), 507–510.
Hardcastle, V. G., & Stewart, M. C. (2002). What do brain data really show? *Philosophy of Science, 69*, S72–S82.
Harvey, E. R. (1975). *The inward wits: Psychological theory in the middle ages and the renaissance*. London: Warburg Institute.
Hoyer, A. (2010). *Neurotechnologie, philosophie und hirnforschung. Zur Entstehung und Institutionalisierung der Neuroethik*. M. A. thesis in philosophy, Johann Wolfgang Goethe-Universität, Frankfurt am Main.
Illes, J. (2003). Neuroethics in a new era of neuroimaging. *American Journal of Neuroradiology, 24*, 1739–1741.
Illes, J. (Ed.). (2006). *Neuroethics: Defining the issues in theory, practice, and policy*. New York: Oxford University Press.
Illes, J., & Sahakian, B. (Eds.). (2011). *Oxford handbook of neuroethics*. New York: Oxford University Press.
Illes, J., Kirschen, M., & Gabrieli, J. (2003). From neuroimaging to neuroethics. *Nature Neuroscience, 6*(3), 205.
Illes, J., & Racine, E. (2005). Imaging or imagining? A neuroethics challenge informed by genetics. *American Journal of Bioethics, 5*(2), 5–18.
Illes, J., Racine, E., & Kirschen, M. P. (2006). A picture is worth a thousand word, but which one thousand? (In Illes 2006, pp. 149–168).
Insel, T. (2010). Rethinking schizophrenia. *Nature, 468*, 187–193.
John, S. (2004). Titanic ethics, pirate ethics, bioethics. *Studies in History and Philosophy of the Biological and Biomedical Sciences, 35*, 177–184. (Essay review of Frankel Paul, E., Miller, F. D. Jr., & Paul J. (Eds.), *Bioethics*. New York: Cambridge University Press, 2002).
Joyce, K. A. (2008). *Magnetic appeal: MRI and the myth of transparency*. Ithaca: Cornell University Press.
Kemp, S. (1990). *Medieval psychology*. New York: Greenwood Press.
Kempton, M., et al. (2008). Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Archives of General Psychiatry, 65*(9), 1017–1032.
Kendler, K. S., & First, M. B. (2010). Alternative futures for the DSM revision process: Iteration v. paradigm shift. *The British Journal of Psychiatry, 197*, 263–265.
Koolschijn, C., et al. (2009). Brain volume abnormalities in major depressive disorder: A meta-analysis of magnetic resonance imaging studies. *Human Brain Mapping, 30*(11), 3719–3735.
Kosslyn, S. M. (1999). If neuroimaging is the answer, what is the question? *Philosophical Transactions of the Royal Society of London, 354*, 1283–1294.
Kulynych, J. (2002). Legal and ethical issues in neuroimaging research: Human subjects protection, medical privacy, and the public communication of research result. *Brain and Cognition, 50*, 345–357. (Also in Glannon 2007, pp. 115–133).
Larink, W. (2011). *Bilder vom Gehirn. Bildwissenschaftliche Zugänge zum Gehirn als Seelenorgan*. Berlin: Akademie Verlag.
Levy, N. (2008). Introducting *neuroethics*. *Neuroethics, 1*, 1–18.
Littlefield, M. M., & Johnson, J. M. (Eds.). (2012). *The neuroscientific turn: Transdisciplinarity in the age of the brain*. Ann Arbor: University of Michigan Press.
Malhi, G. S., & Lagopoulos, J. (2008). Making sense of neuroimaging in psychiatry. *Acta Psychiatrica Scandinavica, 117*, 100–117.
Mandressi, R. (2011). Le cerveau et ses représentations dans la première modernité (XVI^{e}– XVII^{e} siècles). *Médecine/Sciences, 27*, 89–93.

Molnár, Z. (2004). Thomas Willis, the founder of clinical neuroscience. *Nature Reviews Neuroscience, 5*(4), 329–335.
Morris, Z., et al. (2009). Incidental findings on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ, 339*, b3016.
O'Connor, C., & Joffe, H. (2013). How has neuroscience affected lay understanding of personhood? A review of the evidence. *Public Understanding of Science, 22*, 254–268.
Ortega, F. (2011). Toward a genealogy of neuroascesis. In F. Ortega & F. Vidal (Eds.), *Neurocultures: Glimpses into an expanding universe* (pp. 27–44). New York: Peter Lang.
Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man studied by electrical stimulation. *Brain, 60*, 389–443.
Penfield, W., & Rasmussen, T. (1950). *The cerebral cortex of man: A clinical study of localization of function*. New York: Macmillan.
Pogliano, C. (2012). Penfield's *homunculus* and other grotesque creatures from the *Land of If*. *Nuncius, 27*, 141–162.
Poldrack, R. A. (2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed? *Perspectives on Psychological Science, 5*(6), 753–761.
Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews Neuroscience, 6*(2), 159–164.
Racine, E., Bell, E., & Illes, J. (2010a). Can we read minds: Ethical challenges and responsibilities in the use of neuroimaging research. (In Giordano and Gordijn 2010, pp. 244–270).
Racine, E., Waldman, S., Rosenberg, J., & Illes, J. (2010b). Contemporary neuroscience in the media. *Social Science & Medicine, 71*, 725–733.
Rafter, N. (2008). *The criminal brain: Understanding biological theories of crime*. New York: New York University Press.
Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*, 187–206.
Renneville, M. (2000). *Le langage des crânes. Une histoire de la phrénologie*. Paris: Les Empêcheurs de tourner en rond.
Roskies, A. (2002). Neuroethics for the new millenium. *Neuron, 35*, 21–23.
Roskies, A. L. (2007). Are neuroimages like photographs of the brain? *Philosophy of Science, 74*, 860–872.
Roskies, A. L. (2009). Brain-mind and structure-function relationships: A methodological response to coltheart. *Philosophy of Science, 76*, 927–939.
Roskies, A. L. (2010). Saving subtraction: A reply to Van Orden and Paap. *The British Journal for the Philosophy of Science, 61*, 635–665.
Rousseau, G. (2007). "Brainomania": Brain, mind and soul in the long eighteenth century. *British Journal for Eighteenth-Century Studies, 30*, 161–191.
Schick, A. (2005). Neuro exceptionalism? *The American Journal of Bioethics, 5*(2), 36–38.
Schleim, S., & Roiser, J. P. (2009). fMRI in translation: The challenges facing real-world applications. *Frontiers in Human Neuroscience, 3*, Article 63, 1–7.
Shorvon, W. D. (2009). A history of neuroimaging in epilepsy 1909–2009. *Epilepsia, 50*(Suppl), 39–49.
Simpson, J. R. (2012a). Introduction. (In Simpson 2012a, pp. xv–xvii).
Simpson, J. R. (Ed.). (2012b). *Neuroimaging in forensic psychiatry: From the clinic to the courtroom*. Oxford: Wiley-Blackwell.
Singh, I., & Rose, N. (2006). Neuro-forum: An introduction. *BioSocieties, 1*, 97–102.
Spranger, T. M. (Ed.). (2012). *International neurolaw: A comparative analysis*. Berlin: Springer.
Star, S. L. (1989). *Regions of the mind: Brain research and the quest for scientific certainty*. Stanford: Stanford University Press.
Temkin, O. (1973). *Galenism: Rise and decline of a medical philosophy*. Ithaca: Cornell University Press.

Tovino, S. (2007a). Functional neuroimaging information: A case for neuro exceptionalism? *Florida State University Law Review, 34*, 414–489.

Tovino, S. (2007b). Imaging body structure and mapping brain function: A historical approach. *American Journal of Law & Medicine, 333*, 193–228.

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.

Van Orden, G. C., & Paap, K. R. (1997). Functional neuroimages fail to discover pieces of the mind in the parts of the brain. *Philosophy of Science*, *64* (Proceedings), S85–S94.

Van Wyhe, J. (2002). The authority of human nature: the Schädellehre of Franz Joseph Gall. *British Journal for the History of Science, 35*, 17–42.

Vidal, F. (2009). Brainhood, anthropological figure of modernity. *History of the Human Sciences, 22*, 5–36.

Vidal, F. (2011). *The Sciences of the Soul: the Early Modern Origins of Psychology* (trans. Brown, S.). Chicago: University of Chicago Press.

Vidal, F., & Ortega, F. (2011). Are there neural correlates of depression? In S. Choudhury & J. Slaby (Eds.), *Critical neuroscience: A handbook of the social and cultural contexts of neuroscience* (pp. 345–366). Oxford: Wiley-Blackwell.

Willis, T. (1664). *The anatomy of the brain and nerves* (trans: Pordage, S., London, 1681). Birmingham, AL: The Classics of Neurology & Neurosurgery Library, 1983.

Wolf, S. M. (2011). Incidental findings in neuroscience research: A fundamental challenge to the structure of bioethics and health law. (In Illes and Sahakian 2011, pp. 623–634).

Wolpe, P. R. (2002). The neuroscience revolution. *The Hastings Center Report, 32*(4), 8.

Young, R. (1990). *Mind, brain and adaptation in the nineteenth century: Cerebral localization and its biological context from Gall to Ferrier*. New York: Oxford University Press.

Zago, S., Lorusso, L., Ferrucci, R., & Priori, A. (2012). Functional neuroimaging: A historical perspective. In P. Bright (Ed.), *Neuroimaging – Methods* (InTech), Ch. 1. http://www.intechopen.com/books/neuroimaging-methods/the-origins-of-functional-neuroimaging-techniques

Zimmer, C. (2004). *Soul made flesh: The English civil war and the mapping of the mind*. London: Heinemann.

Section VII

Ethical Implications of Brain Stimulation

Ethical Implications of Brain Stimulation

34

Matthis Synofzik

Contents

Abstract

Deep brain stimulation (DBS) is an implanted device allowing to modulate selected brain target sites in individuals' daily life. While it was first tested in the treatment of chronic pain, the main breakthrough was achieved in the treatment of movement disorders. The field of DBS now witnesses an exponential growth in clinical applications and investigations, including diverse psychiatric disorders, behavioral states, neurological orphan diseases, and even enhancement purposes. This leads to fundamental questions on the level of research, clinical and regulatory ethics, which are addressed in this section of the *Handbook of Neuroethics*. This introductory chapter provides an overview of the indications of DBS and the central ethical issues of the burgeoning field.

M. Synofzik
Department of Neurology, Centre for Neurology and Hertie-Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany
e-mail: matthis.synofzik@uni-tuebingen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_28,

Introduction

The idea to directly intervene in the brain and specifically modulate the activity of certain brain circuits by a simple button press seemed to be utopian for long. It has now become partly real by the rapidly increasing development and application of deep brain stimulation (DBS). DBS systems use a chest- or abdomen-implanted pulse generator to send electrical pulses via an extension wire running at the side of the neck to an intracerebral lead which contains several electrodes at the brain target site. Chronic stimulation of subcortical structures was first used already in the early 1950s (Hariz et al. 2010). Since this time, ablative surgery and electrical stimulation developed in parallel, with first applications of these procedures in the realm of psychiatry and behavior (and not in movement disorders, as often assumed) (Hariz et al. 2010). The notion of "deep brain stimulation" was coined in the mid-1970s, at that time investigated as a treatment for chronic pain (Hariz et al. 2010, 2013). The lack of clear indications, thoroughly controlled trials, and longer term follow-up led to the fact that DBS was never approved for pain by the U.S. Federal Drug Administration (FDA) (Hariz et al. 2013). In 1987, however, the group of Benabid published a seminal work on chronic electrical stimulation of the thalamus for the treatment of tremor (Benabid et al. 1987). Since this time, and in particular in the last 15 years, the field of DBS has witnessed an exponential growth in clinical applications and investigations (Hariz et al. 2013). This rapid expansion now includes DBS investigations in diverse psychiatric disorders, behavioral states, and neurological orphan diseases (Table 34.1), and partly builds on only serendipitous and controversial findings and uncontrolled small case studies. This leads to fundamental questions on the level of research, clinical and regulatory ethics, which are addressed in this section of the Handbook of Neuroethics. The purpose of this introductory chapter is to provide an overview of the indications of DBS, the central ethical issues, and the main perspectives of the following sub-chapters.

DBS in Movement Disorders

In the two decades following the seminal work of Benabid's group on DBS for tremor in 1987 (Benabid et al. 1987), DBS has become a well-established and FDA-approved treatment for three classes of movement disorders: Parkinson's disease (PD), primary dystonia, and essential tremor. In these indications, DBS targets various areas in the basal ganglia and thalamus, including the subthalamic nucleus (STN), the globus pallidus internus (GPi), or different parts of the thalamus (e.g., the ventro-intermediate nucleus [VIM]) (Hariz et al. 2013). Although movement disorders and especially PD will certainly remain the main indication for DBS for the foreseeable future, there are still unsettled issues (Clausen 2010). For example, it is still controversially debated if and to what extent the DBS-induced improvements in the motor domain are also paralleled by an overall improvement in quality of life (Agid et al. 2006; Schupbach et al. 2006). Moreover, the psychiatric effects of DBS in

Table 34.1 Established, experimental and theoretical indications for deep brain stimulation (Adapted, modified and extended from (Hariz et al. 2013))

	Established	Experimental	Theoretical
Neurology			
Movement disorders			
Parkinson's disease	x		
Essential tremor and other tremors (cerebellar, post-traumatic, rubral, multiple sclerosis, etc.)	x		
Primary generalized or segmental dystonia	x		
Huntington's disease		x	
Chorea acanthocytosis and other non-Huntington chorea syndromes		x	
Rare hereditary dystonia syndromes (myoclonus dystonia, dystonia-parkinsonism syndromes)		x	
Ataxia syndromes		x	
Wilson's disease		x	
Lance-Adams syndrome		x	
Pain			
Chronic pain		x	
Cluster headache			x
Epilepsy			
Partial epilepsy		x	
Generalized epilepsy		x	
Psychiatry/behavior/cognition			
Gilles de la Tourette syndrome		x	
Obsessive compulsive disorder		x	
Major depression		x	
Bipolar disorder		x	
Schizophrenia			x
Autism		x	
Lesch-Nyhan syndrome and other self-injurious behaviors		x	
Anorexia nervosa		x	
Obesity		x	
Drug/alcohol/nicotine addiction		x	
Cognition			
Alzheimer's dementia		x	
Cognitive impairment in PD		x	
Minimally conscious state (MCS)		x	
Health/enhancement			
Enhancement of memory			x
Enhancement of mood/happiness			x

PD and the effects upon selfhood still warrant a thorough discussion, as outlined in the contribution by ***Paul Foley***. Other authors have shown that DBS ethics does not require any specific ethical criteria, but should follow the widely established bioethical criteria of beneficence, non-maleficence, respect for autonomy and distributive justice (Clausen 2009, 2010; Synofzik and Schlaepfer 2008, 2010, 2011). Foley now applies these criteria to DBS for PD, which serves as a paradigmatic disease for the neuroethical analysis of DBS in movement disorders.

DBS in Psychiatric Diseases

In contrast to widespread assumptions, the renaissance of DBS in surgical treatment of psychiatric illness in 1999 had little to do with nonmotor effects of subthalamic nucleus DBS in PD or with the success of DBS for movement disorders in general (Hariz et al. 2010). Instead, high-frequency DBS for psychiatric diseases took advantage of the experiences of previous psychiatric ablative surgery, in particular using the same brain targets (Hariz et al. 2010). For example, the first modern reports of DBS for obsessive compulsive disorder (OCD) or Gilles de la Tourette syndrome (GTS) aimed at stimulating those targets previously subjected to ablative surgery: the anterior capsule for OCD and medial thalamic structures for GTS (Hariz et al. 2013).

DBS is now (re-)introduced as a treatment modality for various psychiatric and behavioral conditions which not only include OCD and GTS, but also major depression, eating disorders (anorexia nervosa and obesity), aggressiveness/violence, substance addiction, or Alzheimer's dementia. The variety of DBS brain targets tested for each of these diseases is much larger than in DBS for movement disorders. For example, DBS has been tried in nine brain targets for GTS, eight for OCD, and ten for depression (Hariz et al. 2013). The rationale for choosing one particular brain target, but not another, is not always completely coherent. It results from serendipity, theoretical models of the respective brain circuits, brain imaging data (either functional or tractography), previous ablative surgery results, surgeons' and/or psychiatrists' preferences, or on various combinations of the above (Hariz et al. 2013).

It has been cautioned not to prematurely follow an implicit dualism between psychiatric and neurological diseases in the discussion of DBS ethics (Synofzik and Clausen 2011). There are no a priori ethical differences between psychiatric and neurological diseases, and both types of diseases should be analyzed by means of the same bioethical criteria and claims about alleged ethical implications of DBS need to be corroborated by clear empirical evidence in both types of diseases and for each disease specifically. Thus, statements about presumed ethical implications of DBS for a certain psychiatric disease (as well as for a certain neurological disease) require a detailed empirical inquiry. For example, it has to be tested for each disease, whether the respective subject group has sufficient understanding of the risks associated with DBS, has realistic expectations of the likelihood of personal benefit, and can appropriately distinguish between DBS research and routine clinical care (Dunn et al. 2011; Synofzik and Clausen 2011). ***Paul P. Christopher*** and ***Laura B. Dunn*** show that the vast majority of prospective DBS participants with treatment-resistant

depression do have adequate decisional capacity, yet many individuals fail to recognize key difference between DBS research and clinical treatment, a phenomenon known as therapeutic misconception. This phenomenon is probably not restricted to patients with depression, but is likely true for many subjects and diseases where DBS is investigated still on a research basis. In fact, the therapeutic misconception might apply even to surrogate decision-makers as well. Their decision is required in DBS for subjects with Alzheimer's Disease who are increasingly impaired in their decisional capacity, but nevertheless enrolled in DBS trials.

Ethics of DBS Regulation

None of the DBS procedures for neuropsychiatric diseases can already be considered to be "established" (Hariz et al. 2013). In fact, the FDA's recent approval of DBS for OCD as a Humanitarian Device Exemption (HDE) has even been criticized as it removes the requirement for a clinical trial of appropriate size and statistical power (Fins et al. 2011). In contrast to the conventional Investigational Device Exemptions (IDE), where both safety and efficacy have to be demonstrated via properly powered clinical trials, HDEs operate via a lower standard of evidence for safety and do not require strict clinical trials for demonstrating efficacy for approval. Thus, as noted by J. Fins and other prominent colleagues in the field of DBS (Fins et al. 2011), by bypassing the rigors of such trials, premature HDEs for neuropsychiatric DBS put patients at risk, limit opportunities for scientific discovery, and give device manufacturers unique marketing opportunities (Fins et al. 2011). As argued by ***Joseph J. Fins***, the HDE for DBS in OCD is an example of regulatory insufficiency and of the failure to distinguish research from therapy due to the aforementioned therapeutic misconception. He points out that the lower standards of HDEs are meant to facilitate approval of devices that are intended for orphan markets where the disease condition annually affects 4,000 or fewer patients. Since numbers are even lower in rare movement disorders, the same regulatory problem will likely arise not only in the field of neuropsychiatric DBS, but also in the rapidly increasing field of DBS for rare movement disorders like Huntington's Disease, chorea acanthocytosis, dystonia-parkinsonism syndromes, or neurodegeneration with brain iron accumulation (NBIA) (Table 34.1).

Joseph J. Fins demonstrates that DBS devices cannot simply be regulated by methods derived from the regulation of drugs, without a thorough appreciation that devices are different from drugs. Instead, devices require their own ethical and regulatory framework. This framework has to take into account also the context of their use as well as the fact that the manner of a device's approval has implications for therapy and research.

DBS Research Ethics

Serendipitous discoveries and advances in neuroimaging are providing abundant new DBS brain targets for a rapidly increasing number of diverse neurological,

psychiatric, behavioral, and cognitive conditions (Hariz et al. 2013). For example, a case report of DBS of the lateral hypothalamus in an obese patient showed no effect on obesity, but provoked in that patient memory flashbacks (Hamani et al. 2008), which now sparked a trial of DBS for Alzheimer's disease (Laxton et al. 2010). But what happened to all the other single-case DBS interventions for which the primary outcomes were not achieved and no interesting secondary effects were observed? (Schlaepfer and Fins 2010) And is it really justified to launch a cohort study in patients with a severe inherently progressive neurodegenerative disease where subjects will ultimately lose all their cognitive capacities and mobility just based on the "side-effects" of DBS in a subject with a completely different condition?

Many findings in the field of neuropsychiatric DBS (though not in the field of DBS for PD, essential tremor or primary dystonia) rely on single cases or small cohorts, variable methodologies, and differing outcome measures. As argued by ***Matthis Synofzik***, these problems make this field particularly prone to bias and selective reporting, evoke ethical concerns regarding possibly premature expansions to new conditions without appropriate justification and research, and indicate the possibility that media, the public, and institutional review boards might be easily misguided by some reports. He suggests three approaches how these problems might be reduced: by an optimization of trial designs, the implementation of standards of reporting, and the creation of a DBS study register which includes in particular single-case studies or case series.

Intrinsic Ethical Objections Against DBS for Neuropsychiatric Diseases and Enhancement Purposes: Are They Valid?

DBS for neuropsychiatric diseases has been confronted with several ethical concerns which mainly focus on the past uses of psychosurgery, the invasiveness of DBS, and the risks for autonomy and identity. These concerns lead to intrinsic objections against DBS and are deeply rooted in many people's intuitions and frame current ethical discussions. ***Anna Pacholczyk*** scrutinizes these objections, demonstrating that they are mostly not convincing if examined in more detail. They are mainly "rhetorical evocations" and may confound a transparent and conclusive ethical assessment of neuropsychiatric DBS.

The same is true for intrinsic objections against an enhancement use of DBS. The prospects of enhancing memory states (Hamani et al. 2008) and of inducing happiness by a button press (Synofzik et al. 2012) indicate that DBS might serve as an attractive tool not only in disease, but also in healthy subjects in the future (Synofzik and Schlaepfer 2008). But is it ethically legitimate to use DBS even in healthy subjects and for enhancement purposes? Anna Pacholczyk convincingly argues that, for drawing normative arguments to answer this question, it is not helpful to resort to the treatment-enhancement distinction. Instead, the ethical analysis has to be guided by the benefit-harm ratio of DBS enhancement to a particular individual. This implies that DBS enhancement uses cannot be ethically discounted a priori, but warrant a careful consideration of the exact benefit and harm.

Conclusion and Future Directions

There are no convincing intrinsic ethical objections against the use of DBS in neurological or psychiatric diseases – and not even for enhancement purposes in healthy subjects. However, to justify the spread of DBS to larger subject cohorts and novel indications, more empirical research is warranted to provide detailed information on several crucial ethical issues. In particular, it has to be shown for each disease condition that subjects or their surrogate decision-makers have sufficient understanding of the risks associated with DBS, have realistic expectations of the likelihood of personal benefit, and can appropriately distinguish between DBS research and routine clinical care, thus preventing a therapeutic misconception. Moreover, the expansion of DBS needs to be accompanied by the establishment of a novel regulatory framework, which focuses on devices, not on merely adapting drug regulatory policies. Finally, a thorough DBS research ethics needs to developed and uniformly implemented into research, publication, and institutional policies.

Cross-References

▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
▶ Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine
▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
▶ History of Neuroscience and Neuroethics: Introduction
▶ Human Brain Research and Ethics
▶ Informed Consent and the History of Modern Neurosurgery
▶ Neuroenhancement
▶ Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry
▶ Parkinson's Disease and Movement Disorders – Historical and Ethical Perspectives
▶ Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia

References

Agid, Y., Schupbach, M., Gargiulo, M., Mallet, L., Houeto, J. L., Behar, C., . . . Welter, M. L. (2006). Neurosurgery in Parkinson's disease: The doctor is happy, the patient less so? *Journal of Neural Transmission Supplementum*, 70 (Suppl), 409–414.

Benabid, A. L., Pollak, P., Louveau, A., Henry, S., & de Rougemont, J. (1987). Combined (thalamotomy and stimulation) stereotactic surgery of the VIM thalamic nucleus for bilateral Parkinson disease. *Applied Neurophysiology, 50*(1–6), 344–346.

Clausen, J. (2009). Man, machine and in between. *Nature, 457*(7233), 1080–1081.

Clausen, J. (2010). Ethical brain stimulation – Neuroethics of deep brain stimulation in research and clinical practice. *European Journal of Neuroscience, 32*, 1152–1162.

Dunn, L. B., Holtzheimer, P. E., Hoop, J. G., Mayberg, H., Weiss Roberts, L., & Appelbaum, P. S. (2011). Ethical issues in deep brain stimulation research for treatment-resistant depression: Focus on risk and consent. *AJOB Neuroscience, 2*(1), 29–36.

Fins, J. J., Mayberg, H. S., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., . . . Schlaepfer, T. E. (2011). Misuse of the FDA's humanitarian device exemption in deep brain stimulation for obsessive-compulsive disorder. *Health Affairs (Millwood), 30*(2), 302–311. doi:10.1377/hlthaff.2010.0157.

Hamani, C., McAndrews, M. P., Cohn, M., Oh, M., Zumsteg, D., Shapiro, C. M., . . . Lozano, A. M. (2008). Memory enhancement induced by hypothalamic/fornix deep brain stimulation. *Annals of Neurology, 63*(1), 119–123.

Hariz, M. I., Blomstedt, P., & Zrinzo, L. (2010). Deep brain stimulation between 1947 and 1987: The untold story. *Neurosurgical Focus, 29*(2), E1.

Hariz, M., Blomstedt, P., & Zrinzo, L. (2013). Future of brain stimulation: New targets, new indications, new technology. *Movement Disorders, 28*(13), 1784–1792. doi:10.1002/mds.25665.

Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Zumsteg, D., Wennberg, R., Keren, R., . . . Lozano, A. M. (2010). A phase I trial of deep brain stimulation of memory circuits in Alzheimer's disease. *Annals of Neurology, 68*(4), 521–534.

Schlaepfer, T. E., & Fins, J. J. (2010). Deep brain stimulation and the neuroethics of responsible publishing: When one is not enough. *Jama, 303*(8), 775–776.

Schupbach, M., Gargiulo, M., Welter, M. L., Mallet, L., Behar, C., Houeto, J. L., . . . Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*(12), 1811–1816.

Synofzik, M., & Clausen, J. (2011). The ethical differences between psychiatric and neurologic DBS: Smaller than we think? *AJOB Neuroscience, 2*(1), 37–39.

Synofzik, M., & Schlaepfer, T. E. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*(12), 1511–1520.

Synofzik, M., & Schlaepfer, T. E. (2010). Neuromodulation – ECT, rTMS, DBS. In H. Helmchen & N. Sartorius (Eds.), *Ethics in psychiatry. European contributions* (pp. 299–320). Heidelberg: Springer.

Synofzik, M., & Schlaepfer, T. E. (2011). Electrodes in the brain-Ethical criteria for research and treatment with deep brain stimulation for neuropsychiatric disorders. *Brain Stimulation, 4*(1), 7–16.

Synofzik, M., Schlaepfer, T. E., & Fins, J. J. (2012). How happy is too happy? Euphoria, neuroethics and deep brain stimulation of the nucleus accumbens. *AJOB Neuroscience, 2*(1), 37–39.

Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects

35

Paul Foley

Contents

Abstract

The introduction of deep brain stimulation for movement disorders has raised concerns about the impact of therapy upon the personal identity of the patients. It is argued that, while caution is naturally advisable, these concerns have been somewhat exaggerated, and that the ethical approach to the therapy of movement aims to maximize both the motor and psychiatric benefits for the patient, as it is their personal welfare that must be placed at the center of questions of therapy and investigation. The historical development of DBS techniques as applied to the therapy of Parkinson's disease is explored, including their status as less destructive and partially reversible successors to surgical techniques employed in the pre-L-DOPA era.

P. Foley
Unit for History and Philosophy of Science, University of Sydney, Sydney, NSW, Australia

Neuroscience Research Australia, Randwick, Sydney, NSW, Australia
e-mail: p.foley@neura.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_159,

Introduction: Historical Development

Neurotransplantation naturally involves neurosurgery, but deep brain stimulation (DBS) represents the more direct successor to the ablative surgical approaches to the relief of Parkinson's disease (PD) developed prior to the introduction of L-DOPA therapy. Unlike the therapies discussed in ▶ Chap. 29, "Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives," the primary target is not modulation of dopaminergic transmission, and the mechanisms underlying its efficacy remain unclear (Benabid et al. 2009; Miocinovic et al. 2013); nevertheless, the major final site of action is presumed to be, as with dopaminergic therapies, the striatum, and DBS alleviates the same symptoms as L-DOPA, and to the same degree (apart from postural instability, where L-DOPA is superior). Hopes that DBS may slow disease progression via neuroprotection of surviving nigral cells (see, for instance, Spieles-Engemann et al. 2010) have not yet been confirmed in patients; there is some evidence that its benefits, like those of DRT, decline with time (Fasano et al. 2010). Although earlier applications of DBS approaches are recorded, including to the therapy of movement disorders (reviewed: Stahnisch 2008; Hariz et al. 2010), these predecessors have had at best very limited influence upon the development of contemporary DBS, which was introduced as a reversible form of neuro-ablation.

Ablative surgical interventions had been undertaken since the early twentieth century, but only where the "*abject condition of the sufferer and his forlorn life-outlook*" (Meyers 1942) justified the radical procedures involved. Until the mid-1950s, ganglionectomy and other lesions of the pyramidal tracts aimed to control tremor and rigidity; various lesions of the cerebral cortex were also tried, but the ensuing functional losses dissuaded against widespread application (reviewed: Cooper 1961). Russell Meyers experimented with various lesions of the basal ganglia, favoring sectioning of pallidofugal fibers, but the fatality rate of this difficult operation was high (greater than 15 %; other surgeons reported fatality rates of up to 41 %); most neurosurgeons preferred a more direct attack upon the pallidum to relieve rigidity (death rate: 7.5 %).

The adoption of stereotactic surgery in the early 1950s permitted more precise procedures and thus further reduced the risks of the intervention. In 1953, Rolf Hassler and T. Riechert introduced thalamotomy into the surgery of PD on the basis that pallidofugal fibers travel to thalamic nuclei, which in turn projected to the motor cortex. By the late 1950s, lesioning of the ventral intermediate/ventrolateral nuclei had displaced pallidotomy as "*the single most effective lesion for producing complete and enduring relief of parkinsonian tremor and rigidity*" (Cooper, cited by Redfern 1989). Although only 15 % of patients were suitable candidates for such interventions, 90 % of this select group could expect relief from tremor and rigidity (but not akinesia), and stereotactic techniques had reduced the rate of serious complications to about 10 %, and mortality to less than 1 %. The often spectacular abolition of tremor achieved with thalamotomy was not always permanent, with reports of recurrence in 4–20 % of patients; this could be reduced by increasing the lesion size, but only at the cost of temporary (25 %) or permanent (2–9 %) adverse

effects, including motor deficits, dystonia, and speech and sensory disturbances. Bilateral thalamotomy was rare, as it was associated with profound neuropsychological side effects.

Lesioning of the subthalamic nucleus (STN) was introduced in the early 1960s, but the unprecedented effectiveness of the newly introduced L-DOPA therapy threatened to render surgery obsolete for all but the most hopeless cases. Thalamotomy was revived in the 1970s for patients with L-DOPA-unresponsive tremor, while Laitinen and colleagues resumed operations on the medial pallidum in 1985 (detailed reviews of history of PD surgery: Redfern 1989; Benabid et al. 2009; for review of surgery for movement disorders: Krauss and Grossman 2002).

Location of the site to be lesioned was difficult prior to the availability of neuroimaging. Deep brain electrophysiological recording was employed to identify the target region or regions to be avoided (such as the sensory thalamus) according to their characteristic firing patterns; an alternative was electrical stimulation, which could also identify pyramidal tracts. It was in this latter mode that the Grenoble neurosurgeon and professor of biophysics Alim-Louis Benabid and Pierre Pollak, exploring the effects of different pulse frequencies, discovered in 1987 that high frequency stimulation ($\geq$100 Hz) immediately and reversibly abolished parkinsonian tremor (see also Benabid et al. 1991, 2009). This phenomenon had been noted in experimental microelectrode investigations in humans as early as the 1950s (Blomstedt and Hariz 2010), but its therapeutic potential had not been exploited for technical reasons. The opportune discovery echoed Penfield's mapping of the cerebral cortex by stimulation of patients' cortices, to which he had access for unrelated medical reasons: Benabid noted that he stumbled upon the tremor-suppressant effect of high frequencies during a thalamotomy simply because he was curious about what would happen if he varied the stimulus frequency. He was, however, prepared for the discovery, commenting that he had mused beforehand over the different problems associated with L-DOPA therapy and the surgical treatment of PD (Williams 2010).

The first implantation of a permanent electrode, in the ventral intermediate nucleus (Vim) of the thalamus, was undertaken on an explicitly experimental basis, with the subject's consent; the patient had already undergone unilateral thalamotomy, and had for some time sought to have the other thalamus lesioned. The results encouraged a pilot project in patients for whom a second thalamotomy was indicated, again with explicit consent, and again with success, leading to its application in other conditions in which thalamotomy was also employed (dystonia, multiple sclerosis tremor). Seven years after the first implant, the authors regarded the effectiveness, safety, sustained benefit, and absence of serious side effects as recommending its adoption in place of surgical ablation, particularly where bilateral thalamotomy was indicated. The final breakthrough was achieved in 1993 by changing the target to another site, less used in ablative surgery because of the difficult access: the subthalamic nucleus (STN) (Benabid et al. 1993, 1996).

Interest in surgical solutions was promoted at this time by advances in technology, such as neuroimaging, but more so by the development of detailed models of basal ganglia organization and function, based upon extensive animal studies,

particularly of the MPTP model of PD. In this system, the STN is the major excitatory nucleus of the basal ganglia, projecting to both segments of the pallidum and substantia nigra, amongst other regions, while itself receiving input from cortical and subcortical regions (contemporary overviews: Bergman et al. 1990; Parent 1990). Dopamine replacement therapy (DRT) reduces firing of this nucleus, and the aim of DBS is to achieve this inhibition with greater precision, allowing the reduction of L-DOPA dosage and hence of medication-related dyskinesias and fluctuations. DBS-STN had been found to reduce parkinsonian symptoms in the MPTP monkey model of the disease prior to its trialing in human patients. Surgery thus no longer consisted of the *destruction* of a *location* or *neural pathway* based upon empirical findings in the human brain, but explicitly involved *modulation* of one component, a *system* of neural circuits defined in recent animal experiments. It should be noted, however, that this rationale facilitated acceptance of DBS, but did not underlie its introduction; as Benabid et al. (2009) frankly conceded with respect to the its mechanism: "*How this could work? We don't know, but 'who cares, it works.'*" Further, the neurosurgery theater was no longer the preserve of neurosurgeons, with neurologists and neurophysiologists increasingly involved in planning and monitoring of interventions; as Hariz (2003) noted, few publications on PD neurosurgery now include a neurosurgeon as the primary author. In any case, the precise mechanisms underlying its benefits remain to be clarified (Amadio and Boulis 2011).

DBS requires two separate operations: the first implants two electrodes (1½ mm diameter: size of STN = c. 6 × 3½ × 5 mm) into each side of the brain, using stereotactic and brain imaging techniques to locate the electrodes; the second installs a pulse generator, similar to a heart pacemaker, usually located subcutaneously in the chest. The degree of stimulation by this generator can be continuously controlled within pre-set bounds by the patient. Alternative targets for DBS in PD include the internal pallidum (generally regarded as somewhat less effective than DBS-STN, but with fewer psychiatric side effects; recent review: Odekerken et al. 2013) and the ventral intermediate thalamus (in tremor-dominant PD).

DBS is both a safer and more easily regulated successor to ablative surgical approaches, and there is little doubt that DBS achieves the wished-for amelioration of major motor symptoms – it is, in fact, one of the few surgical procedures supported by evidence from double-blind, randomized trials (class I evidence). It alleviates the same symptoms as L-DOPA, and to the same degree (apart from postural instability, where L-DOPA is superior). Its major benefits is that it permits reduction of L-DOPA dosage, and thus reduces the incidence of L-DOPA-related dyskinesias, as well as the vascular and gastrointestinal effects of DRT; it also relieves anxiety and improves mood in most patients (Weaver et al. 2009; Nazzaro et al. 2011; Ashkan et al. 2013; Schuepbach et al. 2013).

DBS is currently licensed for the therapy of PD, dystonia, and essential tremor, and in PD only as a last resort for patients with long-standing, L-DOPA-responsive PD for whom pharmacological methods are not providing adequate relief ('Core Assessment Program for Surgical Interventional Therapies in Parkinson's Disease' (CAPSIT-PD); Defer et al. 1999). Morgante et al. (2007) found that

maximally 4.5 % of 641 PD patients he had assessed qualified on this basis for DBS, with most excluded by insufficiently severe symptoms. More recently, an expert conference arrived at consensus criteria for DBS that defined suitable candidates as "*patients with PD without significant active cognitive or psychiatric problems who have medically intractable motor fluctuations*, *intractable tremor*, *or intolerance of medication adverse effects*." It was noted that optimal results might be achieved only after 3–6 months, and that DBS-STN could be complicated in some patients by increased depression, apathy, impulsivity, impaired verbal fluency, and executive dysfunction (Bronstein et al. 2011; see also Müller and Christen 2011; Jankovic and Poewe 2012). Despite these restrictions, at least 80,000 movement disorder patients had undergone DBS implantation by early 2010.

The most important immediate risks associated with DBS are those of neurosurgery: the likelihood of hemorrhage (2–4 %) and surgery-related infection (2–6 %) are regarded as acceptable. The most important specific complications include mislocation or breakage of the leads, potentially requiring a second, corrective operation. The overall complication rate has been estimated to be about 7 %, a decline from the figure of 25 % during the first 2 years of the century, indicating that most problems can be reduced through experience and technological advances. One undesirable aspect of the procedure is the fact that brain tissue is irreparably destroyed by the insertion of the leads, the long-term consequences of which are unknown (see, for instance, Burdick et al. 2011). While dementia following DBS for PD has been reported, most authors have detected no evidence of this in their patients; as the prevalence of dementia in PD patients in general is as high as 75 % (Jellinger 2012), the consensus is generally that its development probably reflects disease progression rather than a therapy effect (Amadio and Boulis 2011; Baláž et al. 2011b; Clausen 2011; Williams et al. 2011; Weaver et al. 2012). Voges et al. (2007) found that 0.4–1 % of interventions were followed by death or incapacity within the month.

Psychiatric Effects of DBS for Parkinson's Disease

Although it was not discussed in specifically neuroethical terms until 2011, the non-motor effects of DBS in PD have been recognized for more than a decade, and the discussion of the ethics of DBS predates its application to PD (Hariz 2012). The major ethical issue has concerned potential alterations in personality or identity. As these changes are not observed with DBS for essential tremor (a common, pure motor disorder) and dystonia, as stimulation is applied elsewhere (ventromedial thalamus and pallidum, respectively; for the latter, the patient experience has specifically been described as "*being the same inside*, *yet with a new body*": Hariz et al. 2011), so that they probably reflect the fact that the STN serves "*as a nexus that integrates the motor*, *cognitive*, *and emotional components of behavior*", probably as an indirect modulator of both cortical and subcortical structures (Baláž et al. 2011b; see also Mallet et al. 2007).

The effects of DBS upon mood and personality not only raise of variety of ethical questions for PD therapy, but have partially motivated the use of DBS in a variety of non-motor conditions, including obsessive-compulsive disorders, intransigent depression, addiction, and conduct disorders (reviewed: Holtzheimer and Mayberg 2011; Lyons 2011; Wichmann and Delong 2011; Miocinovic et al. 2013), as well as in consciousness disorders (such as vegetative and minimally conscious states: Sen et al. 2010; Giacino et al. 2012; Yamamoto et al. 2013). In these cases, the collateral mental effects reported in PD therapy are among the major desired responses to treatment, and neuroethical concerns regarding DBS for PD were primarily prompted by this possible expansion of use beyond purely neurologic cases. DBS-mediated modification of personality and mood is, in effect, a more refined approach to managing the relevant disorders than psychopharmacology, and probably accompanied by fewer adverse side effects (Synofzik and Schlaepfer 2011; Synofzik et al. 2012). The question of whether use of this technology by healthy persons for purposes of "neural enhancement" is desirable is a separate issue, and only tangentially relevant to movement disorders.

Witt et al. (2008) found that 16.7 % of DBS patients suffered severe psychiatric side effects (mostly apathy and depression) – as did 12.7 % of PD patients receiving best pharmacological treatment, and in this latter group the rate of psychosis was higher. Hälbig (2010), on the other hand, commented that side effects were often reported in uncontrolled studies with small sample sizes, so that the role played by misplaced electrodes or special patient features cannot be assessed. Schüpbach et al. (2006) documented subtle personality changes in unstructured interviews, finding that several patients were more verbose, irritable, and impatient, and that they expressed their opinions more freely; they therefore recommended such interviews for assessing the impact of DBS on personality in order to detect symptoms missed by standardized tests in larger trials (including acute depression, hypomania, and mirthful laughter or pathological crying; see also Krug et al. 2010; Müller et al. 2010; Witt et al. 2011). In their close analysis of personality changes, Müller and Christen (2011) concluded that the small cognitive declines in executive functions, memory, and verbal learning and fluency had only a minor impact on quality of life, although caution was advised in the case of prospective patients already exhibiting problems in these areas. Apathy and depression, both difficult to quantify, were probably related to PD itself, unmasked by reduced dopaminergic medication. Less quantifiable intellectual, affective, and behavioral effects, in contrast, had a potentially major impact on quality of life: complex mental deficits (anticipation and planning, voluntary attention control, organization of complex thoughts) on the one hand, increased energy, novelty-seeking, and risk-taking on the other – none correlated with motor symptom improvement or pre-treatment personality – may clearly have a positive or negative impact upon the patient, depending on their attitude (and that of their environment) to these changes.

There have been occasional reports of impulse control disorders or aggression following DBS-STN (see Frank et al. 2007; Hälbig 2010; Klaming and Haselager 2010; Müller 2010; Broen et al. 2011; Raja and Bentivoglio 2012), but DBS can

also resolve DRT-related behavioral problems, including pathologic gambling, via reduced medication dosage (Müller 2010; Broen et al. 2011); Lhommée et al. (2012) recently proposed that DBS should be specifically offered to DRT patients experiencing addiction problems, although the risk of initial apathy (which they associated with reduced DRT) needs to be carefully monitored in order to protect the patient from depression and suicidal thoughts. It has also been reported that the addiction problem may eventually recur (Shotbolt et al. 2012).

A certain "satisfaction gap" can initially exist between the objective improvement in motor performance achieved by DBS and the subjective state of the patient: "*the doctor is happy*, *the patient less so*" (Agid et al. 2006), "*a distressed mind in a repaired body*" (Schüpbach et al. 2006; see also Krug 2012). This phenomenon requires further exploration, and perhaps dampening of patients' expectations prior to surgery, although even unhappy patients are generally unwilling to discontinue DBS, and most longer term studies find a significantly improved quality of life; a survey of 100 German patients found a median satisfaction rating of 75 %, while only four patients indicated that they regretted the operation (Müller et al. 2010. See also Alvarez et al. 2005; Dörr 2010; Baláž et al. 2011a; Franzini et al. 2011; Smeding et al. 2011). The feelings of the patient immediately after initiation of DBS should not be over-interpreted; a certain initial trepidation and even alienation is natural in someone who now has a metal case in their chest with which they can alter their brain in a manner they do not completely understand; but as they become accustomed to both its benefits and its possible drawbacks, it is viewed as no more exotic then their many pills or their glasses.

Most dissatisfaction, in fact, stems from the personal and social adjustment required by their increased mobility, and from the recognition that even successful DBS does not restore their lives to full health and prosperity. A syndrome termed the "*burden of normality*" may be involved, the biographical adjustment arising from the fact that abrupt amelioration of a severe illness represents a psychological shock that requires time to be managed; the shock may be all the greater than other life events requiring adjustment because mental processing has improved, opening the patient to greater than previous awareness of their situation (Gilbert 2012). Further, the subjective quality of life is not always as reduced by severe neurologic disease as might be supposed; a high degree of adaptation by the patient to their disorder (the degree of acceptance by post-encephalitic parkinsonism patients for as long as half a century was nothing short of astonishing) means that relief of motor symptoms may not ultimately achieve as much for their sense of well-being as anticipated, let alone provide a panacea for all the ills in their lives. Social factors are evidently also involved: Jabre and Bejjani (2007) commented that Lebanese patients did not appear to suffer the same adjustment difficulties as their Western counterparts. A cynic might, indeed, regard post-DBS melancholia as a First World problem: a technological marvel improves a patient's condition so remarkably that they can be either disappointed or overwhelmed by their good fortune. The patient must in any case be counseled prior to surgery with regard to what can and cannot be expected from DBS, although it is unlikely that many prospective candidates will be dissuaded by the fear of disappointment.

Gießen sociologist Helmut Dubiel (2009) provided an inner view of the changes he experienced in response to DBS. Despite much improved motor performance and reduced L-DOPA dosage – which he consciously registered as positive – he felt during the first year post-surgery that he had merely "*replaced the plague with cholera*": depression, impaired olfactory sensitivity, shortness of breath when bending, loss of postural stability, and a shuffling gait that attracted attention in the street were all regarded as being as bad as the banished symptoms, and speech deficits (reduced speed, low volume, slurred articulation) were particularly upsetting for a professor whose self-worth placed great value on his once impressive speaking voice. Dubiel learned that by reducing stimulation he regained his ability to speak clearly – and his accustomed intellectual vitality was also restored – but he was also immobile and depressed; with increased stimulation he could walk comfortably, but not speak. It is less appropriate to discuss such experiences in terms of "identity," as Baylis (2011) has done, than as the natural adjustment to new circumstances that patients must confront during any form of rehabilitation; in fact, Dubiel's insightful account reveals an unshaken identity, a man with an intact mind, but frustrated by the limitations placed upon him by his still defective body.

The rate of attempted suicide is not much higher in DBS patients (1.35 % according to a meta-analysis, one-third realized) than in the general population, but much higher than that for other PD patients, so that DBS patients should be carefully assessed both prior and subsequent to initiation of treatment (Voon et al. 2008; see also Müller 2010). This is particularly important, as suicide is not always preceded by clear depression, and could, as in encephalitis lethargica (EL), be more the consequence of momentary impulsiveness or of a cool assessment of their situation and their future prospects (Müller and Christen 2011; by way of comparison, cf. Fleck 1933 for discussion of rational and impulsive suicide in EL).

A frequently cited extreme example of the psychiatric effects of DBS is the report of Leentjens et al. (2004). A 62-year-old man developed a manic state following successful DBS for advanced PD, in which state he was assessed as being mentally incompetent; reduction in the level of stimulation, however, also restored the motor disability that rendered him bed-bound. The patient was consulted during his DBS-off phase as to whether he wished to continue DBS and be confined in a psychiatric hospital, or to end treatment and be admitted to a nursing home with his intellect intact; the man chose DBS and motor freedom over mental competence. This choice posed several ethical issues, not the least of which was the responsibility of the physicians for averting not only immediate mental harm (mania), but also potential longer term, perhaps irreversible CNS changes: that is, should they (or a legal guardian) override his short-term autonomy to possibly protect his future autonomy (which they might also block)? For a patient of this age and degree of debility, allowing a free decision concerning their own fate is eminently appropriate, as was decided in this instance.

Such cases are, however, highly unusual; more typical is an improvement by DBS-STN of mood and spontaneity. This may itself cause difficulties: the consequently increased enterprise of the patient may disturb their relationship with a spouse or other carer unwilling to surrender their dominant role, or who

conversely expects a greater assumption of responsibility following motor improvement than the patient feels capable. One overview noted that one in eight couples had divorced within two years of DBS (Müller and Christen 2011; see also Schüpbach et al. 2006). Some patients are also unable to continue their previous employment, either because they feel that it is too demanding, or because they prefer to devote themselves to other activities (Schüpbach et al. 2006; Clausen 2011). Relatives may also be alienated by the sense that the changes in their loved one are associated with a man-made device, triggering associations with "cyborgs" (Müller et al. 2010), but in this situation, it is clearly the relatives that require counseling. It has also been found that aspects of *emotional processing* may be modified by DBS-STN (for example, Péron and Dondaine 2012), and this too will modify the interactions of the patient with their environment.

Changes in personality can ultimately be regarded as positive or negative only with respect to how the patient and those closest to them regard them, with respect to how they promote the patient's ability to pursue and achieve personal goals (cf. Synofzik and Schlaepfer 2008; Clausen 2010). Hälbig (2010) cited a case in which a psychiatrically healthy but rather introverted, cautious patient delighted his family with newfound curiosity, spontaneity, and joyfulness (and increased sexual initiative) following DBS; these changes might very well have been interpreted in another relationship as disruptive or unsettling. Even pathological gambling and similar "negative" outcomes may be extreme forms or expressions of the increased spontaneity and reduced inhibition experienced by many patients as a welcome corollary of therapy (cf. Appel-Cresswell and Stoessl 2011; see also Synofzik and Schlaepfer 2008).

There is no doubt that the improvements wrought by DBS represent a major upheaval for all concerned, and Samuel and Brosnan (2011) have criticized the strict focus on the immediate interests of the patient, arguing that such "principlism" (ethical positions based upon the principles of beneficence, non-maleficence, autonomy, and justice) underestimates the significance of social and cultural factors in medical decision-making: "*principlist bioethics focuses narrowly on the patient as an individual*." The detachment from clinical reality evident in this stance would ultimately deny the ability of the patient to make genuinely informed decisions on health matters, given the manifold societal influences upon such decisions, while also warning against any therapy that might result in changes of the relationship of the patient with his surroundings (which would exclude any strategies with an impact upon CNS function or mobility).

It would be hardly ethical to deny the patient treatment in order to satisfy the wishes of their partners: denial of liberation from the captivity imposed by their condition cannot be justified by concerns that they will use their freedom against the wishes of their loved ones. The feelings of the partners should not be dismissed out of hand, particularly as they have often made sacrifices over many years to assist and support their chronically ill loved ones. Robin Mackenzie (2011) asked whether "*family/carers* [*must*] *look after strangers*?", arguing that DBS patients, in contrast to those who have suffered stroke, brain trauma, or dementia, "*may not behave or feel like familiar damaged or diseased loved ones*, *but like healthy strangers with*

claims on family/carers' time, affection, and assets." The personality changes assumed by Mackenzie – they "*lose their capacities for empathy, insight into their own behavior and considering others' interests*" – are from the extreme end of the range encountered in DBS, but his suggestion that consideration of the potential for such changes be incorporated into pre-surgery informed consent documentation has merit. It should nonetheless be recognized that all personalities inevitably change in the course of a relationship – few spouses would argue that their partner was exactly "the same person" as the one they married years or even decades earlier – and that this cannot be interpreted *a priori* as an imposition or affront, particularly where the bone of contention is that the once passive patient had become an active participant in their own life. Haahr et al. (2013) found in their thoughtful article that solidarity characterized a successful relationship before and after DBS, but that the accent shifted from "*living in partnership*" prior to treatment to either "*a sense of freedom embracing life*" or "*the challenge of changes and constraint*" afterwards.

Respecting the autonomy of the patient entails respect for their informed choice of therapy, even where it may result in changes of which a loved one may not approve. Patients and their families should, in any case, be thoroughly advised prior to therapy of the potential for such changes in order to assist both informed consent, as well as adjustment to any alterations that eventuate.

Effects of Deep Brain Stimulation for Parkinson's Disease upon Selfhood

Some commentators maintain that DBS-STN effects perhaps unpermitted modifications of personality or even of identity, that "*DBS arguably tweaks the organ of personhood*" (Amadio and Boulis 2011). There is, however, no a priori reason for a privileged status of the pre-DBS personality that demands its "preservation": not only has the personality of the patient been shaped throughout life by their personal biography, it has also been affected both by the disease process (directly, and in their psychological response to the experience of parkinsonism) and by medication. There is certainly no "genuine" or "underlying personality" that merits restoration, even were this technically feasible (cf. Synofzik and Schlaepfer 2008).

Those who suspect that DBS may alter selfhood typically employ a *relational account* of identity that does not distinguish clearly between "identity" and "personality": "*My identity is not in my body or in my brain, but in the negotiated spaces between my body and brain, and the bodies and brains of others*" (Baylis 2011). The distinction, however, is crucial to understanding people suffering disorders such as PD: as was most dramatically demonstrated by post-encephalitic parkinsonism, motor deficits and bradyphrenia can severely restrict the expressive aspects of personality while permitting the preservation of the underlying identity (Foley 2012).

Facets of *personality* may be tweaked by DBS – but the identity remains untouched. *Identity* is molded to a large degree by interactions with others and with the material world, but is ultimately an internal experience; *personality*

describes elements of this identity as presented to and experienced by others: the relaxed, jovial person valued by a friend is the unserious buffoon decried by his enemies. Personality is more liable to change than identity, as it involves interactions of the person with the outside world, and may be both consciously and unconsciously modified according to context, always in accordance with the preservation of the identity, which is, in contrast, more stable and less amenable to alteration. Robinson Crusoe retained his identity throughout his sojourn, despite the fact that lack of human interaction meant that his personality was no more discernible than light in a vacuum. This distinction between identity and personality was clear to many observers of the EL behavior syndrome during the 1920s, so that many recognized neither "immorality" nor "character change" in the affected children (see ▶ Chap. 29, "Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives") but rather an altered relationship of the physical child with the world imposed by subcortical changes.

Although post-DBS changes have been described as a "*unique form of biographical disruption*" (Gisquet 2008), there is no question of any change so dramatic as to justify description as a "change of identity." Dubiel's insightful account also reveals an unshaken identity, a man with an intact mind but frustrated by the limitations placed upon him by his still defective body, despite the improvements conferred by therapy. As Klaming and Haselager (2010) asked, "*the question that we raise here is whether there is reason to believe that DBS can lead to legally relevant discontinuity type of phenomena regarding memories, intentions, beliefs, goals, desires, and similarity of character between the patient in 'DBS-on' and 'DBS-off' conditions*." The cited case of a 65 year old PD patient who developed manic (disinhibition, amorousness) and psychotic (paranoia) symptoms as an example of personal discontinuity does not amount to a change in identity, unless it is supposed that psychiatric patients in general undergo changes in identity rather than mental function. This direction might be philosophically interesting, but in practical terms is fruitless: mental competence and legal responsibility may be altered in rare cases by DBS-STN, but not the underlying identity.

The fear of external "mind control" and the difference from its psychopharmacological correlate were forcibly expressed by Dubiel (2009) with his observation that "*With long-term medication the neurologic patient becomes a zombie, with a pacemaker he becomes Frankenstein's monster*." This pessimistic view was supported by his neurologist discussing the electronic records of his stimulator, for Dubiel a surprising intrusion into his privacy; he felt that such traces might be used for the social control of his device, and of him. While patient perceptions are important, however, his fears are more a reflection of his unusually intense philosophic analysis of his "implantedness"; most patients instead become accustomed to their device without any definite feelings, positive or negative. It should also be noted that the personality-altering aspects of DBS cannot, in most cases, be turned on and off by means of the stimulator, but are rather the results of middle to longer adjustments following activation of the device.

An unusual case of personality dissociation nevertheless merits attention in this regard: A 43 year old man experienced a dramatic personality dissociation twelve

months after commencement of self-adjusted thalamic DBS for refractory Tourette syndrome. If right side stimulation amplitude was increased, *he* "*anxiously crouching in a corner*, *covering his face with his hands*. *He spoke with a childish high-pitched voice and repeatedly insisted that he was not to blame*. *Sentences were brief and grammatically incorrect*. *If approached by one of us*, *he fiercely kicked his feet because he feared being thrown in the basement*." When stimulation was removed, he was amnestic for this alternate state, remembering only a rush of threatening childhood memories. Assessment of cerebral blood flow supported temporal lobe as well as the fronto-limbic disconnection hypotheses of dissociative disorders (although a direct effect upon the thalamus, long regarded as being involved in the production of a proto-consciousness, should not be excluded) (Goethals et al. 2008). It is important to note that the site of stimulation (thalamus) means that this idiosyncratic response – this is the only case thus far reported – cannot be generalized to DBS in PD. On the other hand, it provides a warning to be heeded in the further application of the technique, particularly as the change in consciousness – if this expression accurately describes the phenomenon observed – was easily manipulated, both of ethical concern with regard to the potential for "mind control", but also encouraging in light of the fact that the "old personality" was neither lost nor permanently impaired.

A hypothetical case in the book *Personality*, *identity and fractured selves* (Mathews et al. 2009) involved a PD patient whose apathy was treated in the near future by DBS. Both the motor symptoms and mood of the patient responded well to therapy, but his "personality" had also changed: Not only was he more outgoing and sociable, he was no longer as devoted to his job as previously; he was more interested in pursuing charitable and political causes, the former Republican was now a solid Democrat. It should be noted, however, that this hypothetical case mixes the known psychological effects of DBS (increased sociability and reduced inhibition) with effects not described in the literature (enhanced intelligence); further, treatment of extreme apathy with DBS is not currently undertaken. Such hypotheticals are interesting intellectual exercises, and may assist to anticipate future developments, but they are of only limited value for addressing current, real situations.

"*The ethical discussion on this point is complicated by the lack of clear and undisputed definitions of central concepts such as personality*, *self*, *identity*, *and authenticity*" (Merkel et al. 2007). The detailed discussions of what constitutes "personality" and "self" are appropriate and important, but are irrelevant to DBS for movement disorders: DBS-STN has no impact upon memory or values (significantly, not even in cases of impulse control), and only exceptionally produces changes that could be detected by standard instruments for the diagnosis of clinical personality disorders. Most instances interpreted as examples of alienation – such as the 38 year old woman with genetic PD (she had suffered PD for 30 years) who complained, "*Now I feel like a machine*, *I've lost my passion*. *I don't recognize myself anymore*" (Schüpbach et al. 2006; see also Kraemer 2011) – involve figures of speech rather than essential changes of identity. Witt et al. (2011) took this as the starting point for a discussion of identity and selfhood, disregarding the fact that the

mental condition described by the patient is entirely comparable with other forms of "fuzziness" or "clouded cognition" – or even simple ennui – that would not normally provoke thoughts of altered identity (cf. also Baylis 2011). The real issue is whether the subject perceives a continuity between their selves before and after DBS, and that they have retained their identity is almost by definition established by their assertion, "I am different to how I was before," "I feel other than I did before." It is more apposite (and banal) to recognize that restoration of control of their own body provides both the physical possibility and the motivation to re-assert their interests in the world: "*I am again myself, and I don't put up with as much as I did before. I stand up to my husband more than I used to*" (Müller et al. 2010). If anything, DBS assists the patient to attain a state that they regard as more authentic and autonomous than previously.

Much of the uneasiness with which DBS is regarded by some is elicited by the implantation of a man-made device into the CNS, an approach that causes more disquietude than the use of less controllable psychopharmacological agents to achieve the same ends. Hälbig (2010), for example, wrote that, "*in parallel with a transfer of his control over mind and body to mechanical devices and to the treating physicians, the patient suffered a personality change with a partial loss of his personal autonomy*." Any transfer of control, however, is in favor of the patient: they control the degree of stimulation – they can even turn the device off permanently – and thereby have greater hegemony over the STN than prior to implantation; voluntary hegemony rather than the automatic regulation enjoyed by non-parkinsonian persons, but self-control nonetheless. This may, ironically, change somewhat if anticipated advances in DBS are realized, such as *closed-loop DBS*, in which feedback from brain electrical activity automatically adjusts stimulation. It is even envisioned that "*within the next decade, hundreds if not thousands of submicron-sized monitoring/modulating electrodes can be placed wherever needed to restore brain function to normal. The term "neuromodulation" will likely replace deep brain stimulation (DBS) as both neurochemistry and electrical activity are included in the therapeutic modalities*" (Andrews 2010a, b).

This anticipated move from "jamming" the STN to more sophisticated modulation of CNS function promises more targeted therapy for patients, and may obviate concerns that chronic continuous stimulation might alter CNS neural activity in an undesirable manner (see, for example, Glannon 2010), but will also exacerbate fears of technological brain control. Two factors will be crucial for ensuring the ethical deployment of such devices: firstly, the patient must retain the option to adjust the activity of such devices, at least within defined parameters that maintain its safe function, and even to turn the device off; secondly, the device should not be subject to external control (for instance, by wireless communication with a computer; cf. Sliwa and Benoist 2011) without the express consent of the patient, and then only under strictly defined conditions dictated by the patient's condition. This issue will probably play a greater role in the extrapolation of DBS to mood and behavioral disorders than in the therapy of movement disorders, as well as in ethically even more problematic extensions to "neural

enhancement" or the control of criminal behaviors, roles in which psychopharmacological agents have played prominent roles.

Fears that patient autonomy are undermined by DBS are often paired with the concern that the patient might use this same autonomy to adjust the level of stimulation to achieve what they feel to the best setting for them, rather than relying upon the Professor to make this decision for them, opening the door to "wish fulfilment medicine" and "neuro-enhancement". Should one begrudge PD patients the opportunity to tweak their mood in a manner that renders life more comfortable? (see Hälbig 2010). Should it even be permitted to trade off aspects of one's psychiatric health in order to achieve motor freedom, should one so desire, as in the Leentjen case? It is ironic that neurologists are often accused of disregarding the emotional needs of their patients, but a therapy that proves to benefit both the motorium and mood of PD patients is regarded with suspicion. Even the best non-motor effects achieved by DBS-STN are hardly grounds for envy (not to mention the unknown motor effects of DBS in a healthy person); it is unlikely that neuro-enhancing pharmaceuticals will be rendered obsolete by expensive elective neurosurgery in the foreseeable future. It has, in any case, nothing to do with the permissibility of DBS in PD. Meaningful functional enhancements that would justify neurosurgery for a healthy consumer remain nothing more than hypothetical, and have nothing to do with use of DBS in movement disorders, or, indeed in certain psychiatric disorders; is the difference between DBS for schizophrenia – once again, under the strict proviso that an appropriate site for DBS can be rationally identified – and the use of a powerful neuroleptic more than an emotional fear of stepping upon a slippery slope that leads to cyborgs and Daleks?

The patient is the final arbiter of how well "they feel in their skin": "*legally competent patients are not obliged to live up to or agree with the ideas of authenticity or rationality held by their doctors*" (Pacholczyk 2011). Ethicist Karsten Witt (2013) argues that decisions by a patient regarding DBS must also consider the hypothetical, changed person they will be after surgery, but effectively (and realistically) depicted this "role-playing" as essentially involving no more than the question "Do I want to become that person?" (cf. also Synofzik and Schlaepfer 2008; Glannon 2009).

Philosophical discussions of an altered "self" or "identity" are, in the end, largely intellectual exercises with respect to DBS therapy of movement disorders. The crude localization of personality traits in particular brain regions is no longer current in neuropsychiatric thought, let alone the localization of identity, nor is it envisaged that the addition or subtraction of nerve cells or stimulation of specific nerve regions, as practiced in the therapy of movement disorders, is any more likely to alter identity to any greater degree than pharmacological therapy of these disorders. It is difficult to envisage that even a healthy person could maintain a stable identity if it were to be subject to modification by every structural or functional change experienced by that person in the course of their life. His compatriots may have remarked that Phineas Gage was "no longer Gage" after his horrendous accident, but Gage himself was in no doubt that he, in fact, was.

DBS is, in short, a "softer" neurosurgical approach to movement disorders, allowing an effective intervention that can be suspended should this prove necessary. There are still unresolved issues: the psychological effects of DBS, for instance, may not be as easily reversible as turning off the stimulator, and even some of the immediate neural effects may be of greater permanence than is generally supposed (via neural plasticity, altered firing patterns); these enduring changes may be beneficial (as predicted by the desynchronization hypothesis of DBS: Wilson et al. 2011) or deleterious. Finally, with increased confidence instilled by experience, it may be considered ethical to trial DBS in younger patients, in whom less advanced disease may allow greater potential for slowing the course of disease (see recent positive reports by Kahn et al. 2012; Schuepbach et al. 2013).

Deep Brain Stimulation as an Experimental Procedure

Although now a well-established therapeutic approach in the treatment of PD, DBS has not entirely shed its experimental aspects; indeed, the success of the method has rendered it ideal for experimental purposes:

> Other reasons why DBS, especially STN DBS, virtually eliminated ablative surgery are that DBS can be safely performed bilaterally and that DBS as a technique allows safe and ethically acceptable studies of the brain in living humans. Anatomy, physiology, psychology, cognition, and many other features could be studied by recording and stimulating at various electrical parameters, as well as by structural imaging, functional imaging, MEG, and so forth, using the DBS electrode as a tool that can be externalized, adjusted, manipulated, or turned on or off if needed in single- or double-blind fashion.
>
> Hence, the marriage between DBS and the STN has opened virtually limitless avenues for research, observation, and documentation, resulting in STN DBS becoming by far the most published functional neurosurgical procedure. (Hariz 2012)

This certainly blurs the divide between therapy and research, and between basic and applied research, but every medical intervention is, in effect, an experiment, as the physician can never be certain how a particular individual will respond to a medication or surgical intervention, no matter how well established; the value of both casuistic reports and post-licensing follow-up studies of medications reflects the fact that unexpected effects – ranging from unpredicted benefits to fatal outcomes – may occur in only a comparatively small number of patients, or long term effects might be recognized only when a larger patient collective has been reviewed. The permissibility of an intervention in an individual case is thus not determined solely by whether it is experimental or not, but by the degree of evidence supporting the rationale underlying the intervention, and the informed consent of the recipient. The results of animal studies cannot be transposed directly onto the human patient, with thalidomide providing a spectacular example where safety in animal models inspired false confidence with regard to use of the drugs in humans. Finally, there is a widespread misconception that "scientific medicine" provides patients with guaranteed solutions. Medical science, however, is constantly evolving, and the clinical answer

offered today can never be more than the best currently available; trial and error is inherent not only in experimental medicine and basic science, but also in the ongoing analysis of the outcome of existing therapeutic strategies. Finally, it would be remiss of a physician or surgeon not to maximize the profit for both patient and medical knowledge to be realized in the course of a therapeutic intervention but always according to the maxims that the patient not be harmed or their time wasted, and that the supplementary investigations be conceptually justified and not simply a whim or the emission of a pet hypothesis.

The cited Hariz comment might raise specters of the psychosurgery of the 1950s, or of the neurophysiological investigations of the American psychiatrist Robert Heath, as reviewed by Baumeister (2000) and Stahnisch (2008). Both the institutional and legal frameworks in which such investigations are now conducted, however, ensure that the a return to past transgressions against human dignity is highly unlikely. As noted by Schmitz-Luhn et al. (2012), for example, German law requires that "*just as is the case for any therapeutic treatment*, *the clinical trial must be clinically and factually justifiable*, *the patient must have given his or her informed consent to the trial*, *and the medical entity must adhere to the established rules of professional conduct*." The once virtually unfettered freedom of judgment accorded to lone investigators has not been constrained by ethical considerations, but also by law.

This advance with regard to medical ethics is mostly clearly exemplified by considering a passage from a report on an NIH-supported scientific conference on the experimental use of "depth electrodes" (mostly employed to explore cortical function) in the mid-1950s, published in the NIH annual report:

> It looks as though, with the methods devised up to the present time, that such methods are going to become relatively popular whether this is warranted therapeutically at present or not. It looks as if it is important to encourage publication and discussion and not to allow ethical judgments to drive people "underground", i.e. to prevent publications and full exchange between investigators. This field seems to be acquiring a respectability and a set of ethics which are acceptable to most of the medical profession and to most scientific investigators in the field. (Lilly 1957)

Several features underscore the fact that investigators were operating in a drastically different legal and ethical context to that which now exists in Western countries. A research direction at this point could "*acquir*[*e*] *respectability and a set of ethics*" in the same way a gentleman's club established its rules – by internal discussion of what was most convenient and tolerable for members, rather than being grounded in *a priori* accepted basic human rights. Ethics was a matter of custom among the investigators alone, rather than a consensus view reached by discussion of specific issues by a broader range of interested persons. This partly explains not only the unbridled experimentation of Heath, but of other practices that continue to inspire horror, such as frontal lobotomy and the excessive employment of shock therapies, practiced in the main, it should be noted, not by neurologists or neurosurgeons, but by overconfident psychiatrists. The historical trajectory between practices in this period and current practices is twofold: the direct path connects the

evidence-based neurosurgery for PD of the period with more recent DBS approaches; the second comprises the warning provided by the failure of self-regulation on an individual basis to secure the interests of patients in biomedical research when the ambition of researchers blinded them to the fact that human dignity should not be sacrificed to the scientific enterprise, no matter how noble its aspirations.

On the other hand, it would unreasonable to proscribe the pursuit of a promising intervention purely on the basis that its mechanisms were not fully understood; this would, for example, also disallow the use of most surgical anesthetics, as well as the "off-label" use of licensed pharmaceuticals for otherwise untreatable conditions. Caution must naturally be applied where the mode of action of an intervention has not been completely elucidated; but scientific medicine has always been a pragmatic endeavor, whereby effective therapies were developed and adopted on the basis of their apparent efficacy, without always knowing why they were effective. Intervention in the brain demands particularly caution, both because of its complexity and the very incomplete understanding of the basis of this complexity, but advances in the areas under discussion will only be achieved by carefully controlled trial and error. Future observers will perhaps regard early twenty-first century DBS techniques as frighteningly primitive, but more satisfactory procedures will develop only through evolution from these beginnings.

A similar situation is evident in the history of DRT for PD. There was a rational basis (the basal ganglia dopamine deficit) for introducing L-DOPA therapy in 1961, but knowledge of the biochemistry of PD was rudimentary: How, for example, could the striatal dopamine deficit be causally related to the hallmark of parkinsonism, nigral degeneration, given that a nigro-striatal pathway had not yet been described in humans? How did dopamine actually work in the CNS? – that it was a neurotransmitter was not yet universally accepted, and many doyens of catecholamine research rejected a role for catecholamines in behavior altogether. The issue was further complicated shortly afterwards by the dopamine hypothesis of schizophrenia: could "too much" dopamine supplementation cause psychosis? The final breakthrough for the therapy was achieved despite the marked negative side effects of high dosage L-DOPA, and its success remained unbroken by the rapidly growing awareness of dyskinesias and other side effects associated with chronic therapy. It was only after L-DOPA had been established as the "gold standard" of PD therapy that the different dopaminergic neural systems were mapped, basal ganglia circuits defined, and families of dopamine receptors identified, contributing to comprehension of the mechanisms underlying the desired and undesired effects of L-DOPA: whereby experience gained in the therapy of PD interacted with laboratory findings to promote advances in both the fundamental research and clinical spheres. Genuinely evidence-based medicine involves a continuous feedback between theory and practice that permits the refinement of neuroscientific models that underlie both theory and therapy.

Neither DRT nor DBS are perfect in their current configurations, but the benefit afforded patients who would otherwise be condemned to the misery unfamiliar to

most Western neurologists of the L-DOPA era means that each can be employed and further developed with clear conscience.

Overview of General Neuroethics of Deep Brain Stimulation for Parkinson's Disease

Autonomy

The impact of current therapeutic strategies for PD does not threaten the stability or continuity of the patient's identity or autonomy. Changes in decision-making are attributable to the increased spontaneity and freedom of expression achieved by therapy, and as such may have undesirable consequences, but do not reflect an insidious "change of character."

The nature of DBS means that treatment is provided by a limited number of centers, presumably including particularly confident advocates of the normal procedures. There are currently only two major manufacturers of the DBS devices employed in PD, so that the clinical trial material assembled on their websites, together with the clinical leaders of those trials, will exert a greater influence on the opinions of prospective patients (who may not recognize the promotional element of "information pages") and their physicians than is achieved even by insistent pharmaceutical advertising. This influence is amplified when the popular press uncritically adopts the manufacturers' claims regarding treatment breakthroughs (Laryionava et al. 2010b; Fins et al. 2011; Gilbert and Ovadia 2011; Schermer 2011).

Non-Maleficence

The selection of therapy for an individual patient must consider the nature of the therapy, the potential side effects, and the current state of the patient; caution must be especially exercised with respect to the physical, cognitive, and emotional status of the patient before undertaking surgery, including their attitudes to surgery and calculated risk. Minimization of the stresses imposed by post-surgery adaptation by comprehensive discussions of what the surgery can and cannot achieve, as well as the consequences this will have for the patient and their families and other close contacts is essential. Patients with a history of depression, psychiatric disease, or impulse control disorders should be carefully monitored both before and subsequent to surgery.

Beneficence

This aspect is assured by implementing only those therapeutic strategies in the treatment of a patient that both have firm theoretical and experimental bases. Clear consensus decision strategies with respect to treatment options

have been established in both the United States and Europe; a conservative approach, commencing with L-DOPA and proceeding to more invasive techniques (including DBS) only where absolutely necessary, minimizes the risk to patients. DBS recipients also require long term specialist care that exceeds the needs of a cardiac pacemaker recipient, for example; the changes associated with successful treatment, in particular, require support from a variety of medical and social support specialists (Fins 2009; Bell et al. 2011b).

Justice

The major issue is whether the benefit afforded by therapy justifies the costs involved. Estimates of the costs associated with DBS indicate that, while the initial surgery is expensive (tens of thousands of dollars), the long-term costs are lower than those of L-DOPA therapy alone, with one study indicating that the initial costs can be recouped in as little as 2.2 years. In particular, expenditure for DRT is markedly reduced, a welcome parallel to the benefit for the patient of their reduced reliance on these medications (reviewed: Valldeoriola et al. 2007; Bell et al. 2011a; Dams et al. 2011; McIntosh 2011; Toft and Dietrichs 2013).

Concluding Remarks

Although ethical questions certainly arose with each new therapeutic approach for PD since the introduction of L-DOPA in the 1960s, and any strategy that alters dopaminergic transmission or basal ganglia function may have an impact upon mood and mentation, discussion of such issues achieved significant volume only with the establishment of DBS-STN – and even then, as several authors have noted, only since 2010, more than 20 years after the introduction of DBS for PD. By this time, a not inconsiderable professional literature concerning the mental and emotional effects of DBS in PD was already available. There was no lack of awareness of the importance of the problems involved, but they were addressed in the practical manner of those concerned with people seeking help: They sought to elucidate factors that might assist in identifying patients at risk of negative outcomes; the importance of multidisciplinary support both before and after surgery was recognized.

There are two interlocked reasons for the uneasiness sometimes expressed with regard to DBS in PD: Firstly, the implantation of a man-made device in the brain is perceived to be intrinsically more disturbing than the use of pharmacological agents to achieve the same ends (and with more frequent adverse effects), or even stimulation of or neurotransplantation in the spinal cord; and secondly, it is seen as the first step to the development of implants with the specific aim of controlling behavior. Such fears are fuelled by press reports of apparent advances in

"mind-reading" by brain imaging equipment, or of systems that will allow quadriplegics to control a computer "directly with their thoughts." Indeed, attention turned to the ethics of DBS for PD only when extrapolation of the use of DBS to the therapy of psychiatric disorders was mooted, triggering memories of older style psychosurgery. A recent survey of German medical students, for instance, found almost unanimous approval for the use of "brain pacemakers" for the management of PD, much less support for their application to the treatment of alcoholism or depression, and almost total rejection for their use for neuroenhancement (Laryionava et al. 2010a).

While the disquiet surrounding implanted devices is quite understandable and cannot be ignored, nor should the fact be overlooked that this conflation of issues and impressions is largely irrelevant to discussion of the ethics of DBS when applied to movement disorders. DBS-STN may improve mood and mental vigor in of PD patients, but thoughts, memories, and goals – the core features of identity – are untouched. It is, in any case, unexpected that improved mood is viewed with suspicion, rather than as a welcome bonus; as Müller (2010) has argued, DBS for PD should not focus purely on motor symptoms, but also on cognitive, emotional, and social factors.

The central problem for skeptics was captured by Merkel et al. (2007) in their declaration:

> Practically no intervention in the structure or functioning of the human brain can be undertaken in complete certainty that it will not affect mental processes, some of which may come to play a key role in a person's self-concept

It is a declaration, however, that applies to any CNS intervention, from taking aspirin for a headache to a complete cerebral lobe extirpation, and as such offers little guidance for the ethical approach to the therapy of PD. Further, it is already possible to define critical brain regions where one should proceed with extreme caution, and research will broaden understanding of which other areas need to be handled with care. This will proceed, as neuroscience always has, by trial and error, braked but also legitimized by ethical boundaries to what research in the CNS is permissible.

I would plead for a less intensively philosophical approach to identity in considering the permissibility of direct CNS interventions in PD, particularly as gaps in our knowledge regarding the relationship of brain and mind are greater than those regarding the extrapyramidal system. It is clear that mind and movement are no more interwoven than they are in the basal ganglia, and this means that any attempt to address movement disorders must also address the attendant psychiatric aspects that, after fifty years of DRT, are increasingly familiar. But pragmatic ethics are required in the clinic, and they should be focused upon the physical and mental well-being of the patient as judged by the patient themselves. Humanistic medicine cannot be a slave to technology, but nor also to quasi-metaphysical speculation that distracts from assisting patients regain autonomy. In a very real sense, the physician hopes that their patient emerges a new person through their therapy – a freer, happier person, more able to pursue their interests than they once were.

Cross-References

▶ Impact of Brain Interventions on Personal Identity
▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives

References

Agid, Y., Schüpbach, M., Mallet, L., Houeto, J. L., Behar, C., Maltete, D., & Welter, M. L. (2006). Neurosurgery in Parkinson's disease: The doctor is happy, the patient less so? *Journal of Neural Transmission. Supplementum, 70*, 409–414.

Alvarez, L., Macias, R., Lopez, G., Alvarez, E., Pavon, N., Rodriguez-Oroz, M. C., Juncos, J. L., Maragoto, C., Guridi, J., Litvan, I., Tolosa, E. S., Koller, W., Vitek, J., DeLong, M. R., & Obeso, J. A. (2005). Bilateral subthalamotomy in Parkinson's disease: Initial and long-term response. *Brain, 128*, 570–583.

Amadio, J. P., & Boulis, N. M. (2011). Practical considerations in the ethics of parkinsonian deep brain stimulation. *AJOB Neuroscience, 2*, 24–26.

Andrews, R. J. (2010a). Neuromodulation. Advances in the next decade. *Annals of the New York Academy of Sciences, 1199*, 212–220.

Andrews, R. J. (2010b). Neuromodulation. Advances in the next five years. *Annals of the New York Academy of Sciences, 1199*, 204–211.

Appel-Cresswell, S., & Stoessl, A. J. (2011). Ethical issues in the management of Parkinson's disease. In J. Illes & B. J. Sahakian (Eds.), *The Oxford handbook of neuroethics* (pp. 575–600). Oxford: Oxford University Press.

Ashkan, K., Samuel, M., Reddy, P., & Chaudhuri, K. R. (2013). The impact of deep brain stimulation on the nonmotor symptoms of Parkinson's disease. *Journal of Neural Transmission, 120*, 639–642.

Baláž, M., Bocková, M., Bareš, M., Rektorová, I., Dírerová, V., & Rektor, I. (2011a). Kvalita života po hluboké mozkové stimulaci u pacientu s pokrocilou Parkinsonovou nemocí. *Ceska a Slovenska Neurologie a Neurochirurgie, 74*, 564–568.

Baláž, M., Bočková, M., Rektorová, I., & Rektor, I. (2011b). Involvement of the subthalamic nucleus in cognitive functions – A concept. *Journal of the Neurological Sciences, 310*, 96–99.

Baumeister, A. (2000). The Tulane electrical brain stimulation program. A historical case study in medical ethics. *Journal of the History of the Neurosciences, 9*, 262–278.

Baylis, F. (2011). "I am who I am": On the perceived threats to personal identity from deep brain stimulation. *Neuroethics*, 14 pp. doi:10.1007/s12152-011-9137-1.

Bell, E., Maxwell, B., McAndrews, M. P., Sadikot, A., & Racine, E. (2011a). Deep brain stimulation and ethics: Perspectives from a multisite qualitative study of Canadian neurosurgical centers. *World Neurosurgery, 76*, 537–547.

Bell, E., Maxwell, B., McAndrews, M. P., Sadikot, A. F., & Racine, E. (2011b). A review of social and relational aspects of deep brain stimulation in Parkinson's disease informed by healthcare provider experiences. *Parkinson's Disease*, 8 pp. ID 871874.

Benabid, A. L., Pollak, P., Gervason, C., Hoffmann, D., Gao, D. M., Hommel, M., Perret, J. E., & de Rougemont, J. (1991). Long-term suppression of tremor by chronic stimulation of the ventral intermediate thalamic nucleus. *Lancet Neurology, 337*, 403–406.

Benabid, A. L., Pollak, P., Seigneuret, E., Hoffmann, D., Gay, E., & Perret, J. (1993). Chronic VIM thalamic stimulation in Parkinson's disease, essential tremor and extra-pyramidal dyskinesias. *Acta Neurochirurgica. Supplementum, 58*, 39–44.

Benabid, A. L., Pollak, P., Gao, D., Hoffmann, D., Limousin, P., Gay, E., Payen, I., & Benazzouz, A. (1996). Chronic electrical stimulation of the ventralis intermedius nucleus of the thalamus as a treatment of movement disorders. *Journal of Neurosurgery, 84*, 203–214.

Benabid, A. L., Chabardes, S., Torres, N., Piallat, B., Krack, P., Fraix, V., & Pollak, P. (2009). Functional neurosurgery for movement disorders: A historical perspective. *Progress in Brain Research, 175*, 379–391.
Bergman, H., Wichmann, T., & DeLong, M. R. (1990). Reversal of experimental parkinsonism by lesions of the subthalamic nucleus. *Science, 249*, 1436–1438.
Blomstedt, P., & Hariz, M. I. (2010). Deep brain stimulation for movement disorders before DBS for movement disorders. *Parkinsonism & Related Disorders, 16*, 429–433.
Broen, M., Duits, A., Visser-Vandewalle, V., Temel, Y., & Winogrodzka, A. (2011). Impulse control and related disorders in Parkinson's disease patients treated with bilateral subthalamic nucleus stimulation: A review. *Parkinsonism & Related Disorders, 17*, 413–417.
Bronstein, J. M., Tagliati, M., Alterman, R. L., Lozano, A. M., Volkmann, J., Stefani, A., Horak, F. B., Okun, M. S., Foote, K. D., Krack, P., Pahwa, R., Henderson, J. M., Hariz, M. I., Bakay, R. A., Rezai, A., Marks, W. J., Moro, E., Vitek, J. L., Weaver, F. M., Gross, R. E., & DeLong, M. R. (2011). Deep brain stimulation for Parkinson disease. An expert consensus and review of key issues. *Archives of Neurology, 68*, 165–171.
Burdick, A. P., Foote, K. D., Wu, S., Bowers, D., Zeilman, P., Jacobson, C. E., Ward, H. E., & Okun, M. S. (2011). Do patient's [sic] get angrier following STN, GPi, and thalamic deep brain stimulation. *NeuroImage, 54*(Suppl. 1), S227–S232.
Clausen, J. (2010). Ethical brain stimulation – Neuroethics of deep brain stimulation in research and clinical practice. *European Journal of Neuroscience, 32*, 1152–1162.
Clausen, J. (2011). Conceptual and ethical issues with brain-hardware interfaces. *Current Opinion in Psychiatry, 24*, 495–501.
Cooper, I. S. (1961). *Parkinsonism. Its medical and surgical therapy*. Springfield, IL: Charles C. Thomas.
Dams, J., Bornschein, B., Reese, J. P., Conrads-Frank, A., Oertel, W. H., Siebert, U., & Dodel, R. (2011). Modelling the cost effectiveness of treatments for Parkinson's disease: A methodological review. *PharmacoEconomics, 29*, 1025–1049.
Defer, G.-L., Widner, H., Marié, R.-M., Rémy, P., Levivier, M., and the conference participants. (1999). Core assessment program for surgical interventional therapies in Parkinson's disease (CAPSIT-PD). *Movement Disorders, 14*, 572–584.
Dörr, D. (2010). Einfluss der Tiefen Hirnstimulation auf die Lebensqualität von Parkinson-Patienten. In D. Groß, G. Gründer, & V. Simonovic (Eds.) *Akzeptanz, Nutzungsbarrieren* und ethische Implikationen neuer Medizintechnologien. *Die Anwendungsfelder Telemedizin und Inkorporierte Technik. Proceedings-Band* (pp. 133–136). Studien des Aachener Kompetenzzentrums für Wissenschaftsgeschichte, 8. Kassel: Kassel University Press.
Dubiel, H. (2009). *Deep in the brain* (trans. P. Schmitz). New York: Europa Editions.
Fasano, A., Romito, L. M., Daniele, A., Piano, C., Zinno, M., Bentivoglio, A. R., & Albanese, A. (2010). Motor and cognitive outcome in patients with Parkinson's disease 8 years after subthalamic implants. *Brain, 133*, 2664–2676.
Fins, J. J. (2009). Deep brain stimulation, deontology and duty: The moral obligation of non-abandonment at the neural interface. *Journal of Neural Engineering, 6*, 050201 (4 pp.).
Fins, J. J., Schlaepfer, T. E., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., Merkel, R., & Mayberg, H. S. (2011). Ethical guidance for the management of conflicts of interest for researchers, engineers and clinicians engaged in the development of therapeutic deep brain stimulation. *Journal of Neural Engineering, 8*, 033001 (6 pp.).
Fleck, U. (1933). Über Selbstmorde und Selbstmordversuche bei Postencephalitikern mit Bemerkungen über die Persönlichkeitsänderung der erwachsenen parkinsonistischen Postencephalitiker. *Archiv für Psychiatrie und Nervenkrankheiten, 99*, 233–300.
Foley, P. (2012). The encephalitis lethargica patient as a window on the soul. In L. S. Jacyna & S. T. Casper (Eds.), *The neurological patient in history* (pp. 184–211). Rochester, NY: University of Rochester Press.
Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: Impulsivity, deep brain stimulation, and medication in parkinsonism. *Science, 318*, 1309–1312.

Franzini, A., Cordella, R., Messina, G., Marras, C. E., Romito, L. M., Carella, F., Albanese, A., Rizzi, M., Nardocci, N., Zorzi, G., Zekay, E., & Broggi, G. (2011). Deep brain stimulation for movement disorders. Considerations on 276 consecutive patients. *Journal of Neural Transmission, 118*, 1497–1510.

Giacino, J., Fins, J. J., Machado, A., & Schiff, N. D. (2012). Central thalamic deep brain stimulation to promote recovery from chronic posttraumatic minimally conscious state: Challenges and opportunities. *Neuromodulation, 15*, 339–349.

Gilbert, F. (2012). The burden of normality: From 'chronically ill' to 'symptom free'. New ethical challenges for deep brain stimulation postoperative treatment. *Journal of Medical Ethics, 38*, 408–412.

Gilbert, F., & Ovadia, D. (2011). Deep brain stimulation in the media: Over-optimistic portrayals call for a new strategy involving journalists and scientists in ethical debates. *Frontiers in Integrative Neuroscience, 5*, 16 (6 pp.).

Gisquet, E. (2008). Cerebral implants and Parkinson's disease: A unique form of biographical disruption? *Social Science & Medicine, 67*, 1847–1851.

Glannon, W. (2009). Stimulating brains, altering minds. *Journal of Medical Ethics, 35*, 289–292.

Glannon, W. (2010). Consent to deep brain stimulation for neurological and psychiatric disorders. *Journal of Clinical Ethics, 21*, 104–111.

Goethals, I., Jacobs, F., Van der Linden, C., Caemaert, J., & Audenaert, K. (2008). Brain activation associated with deep brain stimulation causing dissociation in a patient with Tourette's syndrome. *Journal of Trauma & Dissociation, 9*, 543–549.

Haahr, A., Kirkevold, M., Hall, E. O. C., & Østergaard, K. (2013). 'Being in it together': Living with a partner receiving deep brain stimulation for advanced Parkinson's disease – A hermeneutic phenomenological study. *Journal of Advanced Nursing, 69*, 338–347.

Hälbig, T. D. (2010). Manipulating the brain. An ethical challenge? Lessons from deep brain stimulation in movement disorders. In H. Fangerau, J. M. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 251–280). Bielefeld: Transcript.

Hariz, M. I. (2003). From functional neurosurgery to "interventional" neurology: Survey of publications on thalamotomy, pallidotomy, and deep brain stimulation for Parkinson's disease from 1966 to 2001. *Movement Disorders, 18*, 845–852.

Hariz, M. (2012). Twenty-five years of deep brain stimulation: Celebrations and apprehensions. *Movement Disorders, 27*, 930–933.

Hariz, M. I., Blomstedt, P., & Zrinzo, L. (2010). Deep brain stimulation between 1947 and 1987: The untold story. *Neurosurgical Focus, 29*, E1 (10 pp.).

Hariz, G.-M., Limousin, P., Tisch, S., Jahanshahi, M., & Fjellman-Wiklund, A. (2011). Patients' perceptions of life shift after deep brain stimulation for primary dystonia – A qualitative study. *Movement Disorders, 26*, 2101–2106.

Hassler, R., & Riechert, T. (1954). Indikationen and Lokalizationensmethode der Hirnoperationen. *Nervenarzt, 25*, 441–447.

Holtzheimer, P. E., & Mayberg, H. S. (2011). Deep brain stimulation for psychiatric disorders. *Annual Review of Neuroscience, 34*, 289–307.

Jabre, M. G., & Bejjani, B.-P. W. (2007). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? [letter]. *Neurology, 68*, 1164.

Jankovic, J., & Poewe, W. (2012). Therapies in Parkinson's disease. *Current Opinion in Neurology, 25*, 433–447.

Jellinger, K. A. (2012). Neurobiology of cognitive impairment in Parkinson's disease. *Expert Review of Neurotherapeutics, 12*, 1451–1466.

Kahn, E., D'Haese, P.-F., Dawant, B., Allen, L., Kao, C., Charles, P. D., & Konrad, P. (2012). Deep brain stimulation in early stage Parkinson's disease: Operative experience from a prospective, randomized clinical trial. *Journal of Neurology, Neurosurgery, and Psychiatry, 83*, 164–170.

Klaming, L., & Haselager, P. (2010). Did my brain implant make me do it? Questions raised by DBS regarding psychological continuity, responsibility for action and mental competence. *Neuroethics* doi:10.1007/s12152-010-9093-1 (13pp).

Kraemer, F. (2011). Me, myself and my brain implant: Deep brain stimulation raises questions of personal authenticity and alienation. *Neuroethics*. doi:10.1007/s12152-011-9115-7 (15pp).

Krauss, J. K., & Grossman, R. (2002). Surgery for Parkinson's disease and hyperkinetivc movement disorders. In J. J. Jankovic & E. Tolosa (Eds.), *Parkinson's disease and movement disorders* (pp. 640–662). Philadelphia: Lippincott Williams & Wilkins.

Krug, H. (2012). Das Unzufriedenheitsparadox in der Therapie mit tiefer Hirnstimulation. *Nervenheilkunde, 31*, 215–219.

Krug, H., Müller, O., & Bittner, U. (2010). Technisierung des Ich? Überlegungen zu einer ethischen Beurteilung der tiefen Hirnstimulation unter Verwendung von Patienten-Narrationen. *Fortschritte der Neurologie Psychiatrie, 78*, 644–651.

Laryionava, K., Kreucher, S., Simonovic, V., & Groß, D. (2010a). Einstellung zur Tiefen Hirnstimulation unter Medizinstudierenden. In D. Groß, G. Gründer, & V. Simonovic (Eds.), *Akzeptanz, Nutzungsbarrieren und ethische Implikationen neuer Medizintechnologien. Die Anwendungsfelder Telemedizin und Inkorporierte Technik. Proceedings-Band* (Studien des Aachener Kompetenzzentrums für Wissenschaftsgeschichte, Vol. 8, pp. 111–116). Kassel: Kassel University Press.

Laryionava, K., Simonovic, V., & Groß, D. (2010b). Tiefe Hirnstimulation im Spiegel der deutschen Laienpresse. In D. Groß, G. Gründer, & V. Simonovic (Eds.), *Akzeptanz, Nutzungsbarrieren und ethische Implikationen neuer Medizintechnologien. Die Anwendungsfelder Telemedizin und Inkorporierte Technik. Proceedings-Band* (Studien des Aachener Kompetenzzentrums für Wissenschaftsgeschichte, Vol. 8, pp. 97–103). Kassel: Kassel University Press.

Leentjens, A. F. G., Visser-Vandewalle, V., Temel, Y., & Herhey, F. R. J. (2004). Manipuleerbare wilsbekwaamheid: een ethisch probleem bij elektrostimulatie van de nucleus subthalamicus voor ernstige ziekte van Parkinson. *Nederlands Tijdschrift voor Geneeskunde, 148*, 1394–1398.

Lhommée, E., Klinger, H., Thobois, S., Schmitt, E., Ardouin, C., Bichon, A., Kistner, A., Fraix, V., Xie, J., Kombo, M. A., Chabardès, S., Seigneuret, E., Benabid, A.-L., Mertens, P., Polo, G., Carnicella, S., Quesada, J.-L., Bosson, J.-L., Broussolle, E., Pollak, P., & Krack, P. (2012). Subthalamic stimulation in Parkinson's disease: Restoring the balance of motivated behaviours. *Brain, 135*, 1463–1477.

Lilly, J. C. (1957). Report on conference on the use of depth electrodes in human atients. In *Annual report of program activities. National Institutes of Health. 1957. National Institute of Mental Health* (pp. 258–259). Washington, DC: National Institutes of Health; Public Health Service; U.S. Department of Health, Education, and Welfare.

Lyons, M. K. (2011). Deep brain stimulation: Current and future clinical applications. *Mayo Clinic Proceedings, 86*, 662–672.

Mackenzie, R. (2011). Must family/carers look after strangers? Post-DBS identity changes and related conflicts of interest. *Frontiers in Integrative Neuroscience, 5*, 12 (2 pp).

Mallet, L., Schüpbach, M., N'Diaye, K., Remy, P., Bardinet, E., Czernecki, V., Welter, M.-L., Pelissolo, A., Ruberg, M., Agid, Y., & Yelnik, J. (2007). Stimulation of subterritories of the subthalamic nucleus reveals its role in the integration of the emotional and motor aspects of behavior. *PNAS, 104*, 10661–10666.

Mathews, D. J. H., Bok, H., & Rabins, P. V. (2009). *Personal identity and fractured selve. Perspectives from philosophy, ethics, and neuroscience*. Baltimore: Johns Hopkins University.

McIntosh, E. S. (2011). Perspective on the economic evaluation of deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*, 19 (7 pp).

Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B., & Rosahl, S. (2007). *Intervening in the brain. Changing psyche and society. Ethics of science and technology assessment* (Vol. 29). Berlin/Heidelberg: Springer.

Meyers, R. (1942). The modification of alternating tremors, rigidity and festination by surgery of the basal ganglia. In T. J. Putnam (Ed.), *The diseases of the basal ganglia. Proceedings of the*

Association, December 20 and 21, 1940, New York (Association for Research in Nervous and Mental Disease, Vol. 21, pp. 602–665). New York: Hafner Publishing Company.

Miocinovic, S., Somayajula, S., Chitnis, S., & Vitek, J. L. (2013). History, applications, and mechanisms of deep brain stimulation. *JAMA Neurology, 70*, 163–171.

Morgante, L., Morgante, F., Moro, E., Epifanio, A., Girlanda, P., Ragonese, P., Antonini, A., Barone, P., Bonuccelli, U., Contarino, M. F., Capus, L., Ceravolo, M. G., Marconi, R., Ceravolo, R., D'Amelio, M., & Savettieri, G. (2007). How many parkinsonian patients are suitable candidates for deep brain stimulation of subthalamic nucleus? Results of a questionnaire. *Parkinsonism & Related Disorders, 13*, 528–531.

Müller, S. (2010). Personality changes through deep brain stimulation of the subthalamic nucleus in parkinsonian patients – An ethical discussion. In H. Fangerau, J. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 223–250). Bielefeld: Transcript.

Müller, S., & Christen, M. (2011). Deep brain stimulation in parkinsonian patients – Ethical evaluation of cognitive, affective, and behavioral sequelae. *AJOB Neuroscience, 2*, 3–13.

Müller, O., Bittner, U., & Krug, H. (2010). Narrative Identität bei Therapie mit "Hirnschrittmacher". Zur Integration von Patienten-Selbstbeschreibungen in die ethische Bewertung der tiefen Hirnstimulation. *Ethik in der Medizin, 22*, 303–315.

Nazzaro, J. M., Pahwa, R., & Lyons, K. E. (2011). The impact of bilateral subthalamic stimulation on non-motor symptoms of Parkinson's disease. *Parkinsonism & Related Disorders, 17*, 606–609.

Odekerken, V. J. J., van Laar, T., Staal, M. J., Mosch, A., Hoffmann, C. F. E., Nijssen, P. C. G., Beute, G. N., van Vugt, J. P. P., Lenders, M. W. P. M., Contarino, M. F., Mink, M. S. J., Bour, L. J., van den Munckhof, P., Schmand, B. A., de Haan, R. J., Schuurman, P. R., & de Bie, R. M. A. (2013). Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): A randomised controlled trial. *Lancet Neurology, 12*, 37–44.

Pacholczyk, A. (2011). DBS makes you feel good! Why some of the ethical objections to the use of DBS for neuropsychiatric disorders and enhancement are not convincing. *Frontiers in Integrative Neuroscience, 5*, 1 (2 pp.).

Parent, A. (1990). Extrinsic connections of the basal ganglia. *Trends in Neuroscience, 13*, 254–258.

Péron, J., & Dondaine, T. (2012). Émotion et noyaux gris centraux (II): Que peut-nous apprendre le modèle de la stimulation cérébrale profonde du noyau subthalamique dans la maladie de Parkinson? *Revue Neurologique, 168*, 642–648.

Raja, M., & Bentivoglio, A. R. (2012). Impulsive and compulsive behaviors during dopamine replacement treatment in Parkinson's disease and other disorders. *Current Drug Safety, 7*, 63–75.

Redfern, R. M. (1989). History of stereotactic surgery for Parkinson's disease. *British Journal of Neurosurgery, 3*, 271–304.

Samuel, G. N., & Brosnan, C. (2011). Deep brain stimulation for Parkinson's disease: A critique of principlism as a framework for the ethical analysis of the decision-making process. *AJOB Neuroscience, 2*, 20–22.

Schermer, M. (2011). Ethical issues in deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*, 17 (5 pp.).

Schmitz-Luhn, B., Katzenmeier, C., & Woopen, C. (2012). Law and ethics of deep brain stimulation. *International Journal of Law and Psychiatry, 35*, 130–136.

Schuepbach, W. M. M., Rau, J., Knudsen, K., Volkmann, J., Krack, P., Timmermann, L., Hälbig, T. D., Hesekamp, H., Navarro, S. M., Meier, N., Falk, D., Mehdorn, M., Paschen, S., Maarouf, M., Barbe, M. T., Fink, G. R., Kupsch, A., Gruber, D., Schneider, G.-H., Seigneuret, E., Kistner, A., Chaynes, P., Ory-Magne, F., Courbon, C. B., Vesper, J., Schnitzler, A., Wojtecki, L., Houeto, J.-L., Bataille, B., Maltête, D., Damier, P., Raoul, S., Sixel-Doering, F., Hellwig, D., Gharabaghi, A., Krüger, R., Pinsker, M. O., Amtage, F., Régis, J.-M., Witjas, T., Thobois, S., Mertens, P., Kloss, M., Hartmann, A., Oertel, W. H., Post, B.,

Speelman, H., Agid, Y., Schade-Brittinger, C., Deuschl, G., & the EARLYSTIM Study Group. (2013). Neurostimulation for Parkinson's disease with early motor complications. *New England Journal of Medicine, 368*, 610–622.

Schüpbach, M., Gargiulo, M., Welter, M. L., Mallet, L., Béhar, C., Houeto, J. L., Maltête, D., Mesnage, V., & Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*, 1811–1816.

Sen, A. N., Campbell, P. G., Yadla, S., Jallo, J., & Sharan, A. D. (2010). Deep brain stimulation in the management of disorders of consciousness: A review of physiology, previous reports, and ethical considerations. *Neurosurgical Focus, 29*, E14 (6 pp.).

Shotbolt, P., Moriarty, J., Costello, A., Jha, A., David, A., Ashkan, K., & Samuel, M. (2012). Relationships between deep brain stimulation and impulse control disorders in Parkinson's disease, with a literature review. *Parkinsonism & Related Disorders, 181*, 10–16.

Sliwa, J., & Benoist, E. (2011). Wireless sensor and actor networks: E-Health, e-Science, e-Decisions. In *International conference on selected topics in mobile and wireless networking, iCOST 2011:* 6085829 (6 pp).

Smeding, H. M. M., Speelman, J. D., Huizenga, H. M., Schuurman, P. R., & Schmand, B. (2011). Predictors of cognitive and psychosocial outcome after STN DBS in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry, 82*, 754–760.

Spieles-Engemann, A. L., Behbehani, M. M., Collier, T. J., Wohlgenant, S. L., Steece-Collier, K., Paumier, K., Daley, B. F., Gombash, S., Madhavan, L., Mandybur, G. T., Lipton, J. W., Terpstra, B. T., & Sortwell, C. E. (2010). Stimulation of the rat subthalamic nucleus is neuroprotective following significant nigral dopamine neuron loss. *Neurobiology of Disease, 39*, 105–115.

Stahnisch, F. (2008). Über Forschungsentwicklungen der Neurostimulation nach 1945: Historische und ethische Aspekte medizinischer Manipulationen am menschlichen Gehirn. *Würzburger medizinhistorische Mitteilungen, 27*, 307–346.

Synofzik, M., & Schlaepfer, T. E. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*, 1511–1520.

Synofzik, M., & Schlaepfer, T. E. (2011). Electrodes in the braind. Ethical criteria for research and treatment with deep brain stimulation for neuropsychiatric disorders. *Brain Stimulation, 4*, 7–16.

Synofzik, M., Schlaepfer, T. E., & Fins, J. J. (2012). How happy is too happy? Euphoria, neuroethics, and deep brain stimulation of the nucleus accumbens. *AJOB Neuroscience, 3*, 30–36.

Toft, M., & Dietrichs, E. (2013, in press). Medication costs following subthalamic nucleus deep brain stimulation for Parkinson's disease [letter]. *Movement Disorders.* doi:10.1002/mds.25504.

Valldeoriola, F., Morsi, O., Tolosa, E., Rumià, J., Martí, M. J., & Martínez-Martín, P. (2007). Prospective comparative study on cost-effectiveness of subthalamic stimulation and best medical treatment in advanced Parkinson's disease. *Movement Disorders, 22*, 2183–2191.

Voges, J., Koulousakis, A., & Sturm, V. (2007). Deep brain stimulation for Parkinson's disease. *Acta Neurochirurgica. Supplement, 97*, 171–184.

Voon, V., Krack, P., Lang, A. E., Lozano, A. M., Dujardin, K., Schüpbach, M., D'Ambrosia, J., Thobois, S., Tamma, F., Herzog, J., Speelman, J. D., Samanta, J., Kubu, C., Rossignol, H., Poon, Y.-Y., Saint-Cyr, J. A., Ardouin, C., & Moro, E. (2008). A multicentre study on suicide outcomes following subthalamic stimulation for Parkinson's disease. *Brain, 131*, 2720–2728.

Weaver, F. M., Follett, K., Stern, M., Hur, K., Harris, C., Marks, W. J., Rothlind, J., Sagher, O., Reda, D., Moy, C. S., Pahwa, R., Burchiel, K., Hogarth, P., Lai, E. C., Duda, J. E., Holloway, K., Samii, A., Horn, S., Bronstein, J., Stoner, G., Heemskerk, J., & Huang, G. D. (2009). Bilateral deep brain stimulation vs best medical therapy for patients with advanced Parkinson disease: A randomized controlled trial. *Journal of the American Medical Association, 301*, 63–73.

Weaver, F. M., Follett, K. A., Stern, M., Luo, P., Harris, C. L., Hur, K., Marks, W. J., Rothlind, J., Sagher, O., Moy, C., Pahwa, R., Burchiel, K., Hogarth, P., Lai, E. C., Duda, J. E., Holloway, K., Samii, A., Horn, S., Bronstein, J. M., Stoner, G., Starr, P. A., Simpson, R., Baltuch, G., De Salles, A., Huang, G. D., Reda, D. J., & CSP 468 Study Group. (2012). Randomized trial of deep brain stimulation for Parkinson disease: Thirty-six-month outcomes. *Neurology, 79*, 55–65.

Wichmann, T., & Delong, M. R. (2011). Deep-brain stimulation for basal ganglia disorders. *Basal Ganglia, 1*, 65–77.

Williams, R. (2010). Alim-Louis Benabid: Stimulation and serendipity. *Lancet Neurology, 9*, 1152.

Williams, A. E., Arzola, G. M., Strutt, A. M., Simpson, R., Jankovic, J., & York, M. K. (2011). Cognitive outcome and reliable change indices two years following bilateral subthalamic nucleus deep brain stimulation. *Parkinsonism & Related Disorders, 17*, 321–327.

Wilson, C. J., Beverlin, B., & Netoff, T. (2011). Chaotic desynchronization as the therapeutic mechanism of deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*, 50 (11 pp.).

Witt, K. (2013). Das Identitätsproblem der tiefen Hirnstimulation und einige seiner praktischen Implikationen. *Ethik in der Medizin, 25*, 5–18.

Witt, K., Daniels, C., Reiff, J., Krack, P., Volkmann, J., Pinsker, M. O., Krause, M., Tronnier, V., Kloss, M., Schnitzler, A., Wojtecki, L., Botzel, K., Danek, A., Hilker, R., Sturm, V., Kupsch, A., Karner, E., & Deuschl, G. (2008). Neuropsychological and psychiatric changes after deep brain stimulation for Parkinson's disease: A randomised, multicentre study. *Lancet Neurology, 7*, 605–614.

Witt, K., Kuhn, J., Timmermann, L., Zurowski, M., & Woopen, C. (2011). Deep brain stimulation and the search for identity. *Neuroethics.* doi:10.1007/s12152-011-9100-1 (13 pp).

Yamamoto, T., Katayama, Y., Obuchi, T., Kobayashi, K., Oshima, H., & Fukaya, C. (2013). Deep brain stimulation and spinal cord stimulation for vegetative state and minimally conscious state. *World Neurosurgery. 80,* S30.e1-e9.

Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia

36

Paul P. Christopher and Laura B. Dunn

Contents

Abstract

Interest in deep brain stimulation (DBS) for treatment-refractory psychiatric disorders has raised ethical concerns about the adequacy of safeguards for human subjects. The nature of DBS and potential vulnerability of participants have raised worries about whether study participants have capacity to give informed consent (including understanding the risks associated with DBS and the likelihood of personal benefit), and can appropriately distinguish between DBS research and clinical care. Empirical inquiry into these concerns suggests that the vast majority of prospective participants under consideration for early trials of DBS for treatment-resistant depression (TRD) had adequate decisional capacity. While these patients considered DBS for various and often unique

P.P. Christopher (✉)
Department of Psychiatry & Human Behavior, Alpert Medical School, Brown University, Providence, RI, USA
e-mail: paul_christopher@brown.edu

L.B. Dunn
Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA
e-mail: Laura.dunn@ucsf.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_31,

reasons, most understood the risks and had reasonable expectations of benefit. These results suggest that given appropriate procedures, patients with TRD can provide informed consent to DBS research. Further attention is needed to understand patients' considerations for DBS research, and to evaluate procedures to enhance appreciation of distinctions between research and clinical contexts. Similar inquiry on DBS research for other psychiatric disorders is needed, particularly because greater cognitive impairment is associated with worse performance on measures of decisional capacity. Safeguards such as surrogate consent should be considered (and their use reported) when studies enroll patients with cognitive disorders, such as Alzheimer's disease and other dementias. Research is needed into whether patients with obsessive-compulsive disorder are more likely than those with depression to lack adequate capacity.

Introduction

Following the successful use of deep brain stimulation (DBS) in ameliorating symptoms associated with severe, refractory Parkinson's disease, essential tremor, and primary dystonia (Lyons 2011; Weaver et al. 2012), and its more recent use in treating less severe forms Parkinson's disease (Schuepbach et al. 2013), there has been burgeoning interest in applying DBS in the treatment of a number of psychiatric illnesses. A growing series of small studies suggest that DBS can be an effective intervention for treatment-resistant depression (TRD), obsessive-compulsive disorder (OCD), and Tourette's syndrome (Bewernick et al. 2010; Flaherty et al. 2005; Goodman et al. 2010; Greenberg et al. 2006; Holtzheimer et al. 2012; Holtzheimer and Mayberg 2011; Kennedy et al. 2011; Lozano et al. 2008, 2012; Malone et al. 2009; Mayberg et al. 2005). More recent research has focused on using DBS for Alzheimer disease, other dementias (Laxton and Lozano 2012), anorexia nervosa (Lipsman et al. 2013), and for substance addiction although, thus far, outcomes data are limited.

From its earliest use, DBS has sparked ethical debate (Siegfried et al. 1980). In anticipation of the research application of DBS to treat psychiatric illness, however, increasing attention and concern has been directed toward a specific set of empirical and normative issues. These include the need to provide adequate protections for human research subjects through robust informed consent with clear articulation of the risks, potential benefits, and alternatives to DBS. There is concern that prospective participants, by virtue of their mental illness, may lack decision-making capacity regarding enrollment or may have misconceptions either about the likely benefit from DBS or the experimental nature of the study. Additional concerns include the precipitous application of DBS for treatment-refractory illnesses in the absence of appropriate justification, and ensuring rational scientific methodology and transparency of researchers' conflicts of interest (Bell et al. 2009; Dunn et al. 2011; Fins et al. 2011b; Rabins et al. 2009). Others have cautioned against selective reporting of DBS research in the scientific literature (Schlaepfer

and Fins 2010) as this may contribute to the misrepresentation of DBS efficacy in the lay media (Racine et al. 2007) and, perhaps more importantly, may prematurely or inaccurately guide both clinical practice and future DBS research.

There has been long-standing concern about whether individuals with serious psychiatric illness are able to make rational, informed decisions regarding research enrollment (Elliott 1997; Michels 2004; National Bioethics Advisory 1998; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978). Yet, numerous empirical assessments of subjects with a range of mental illnesses suggest that a majority have the requisite decision-making capacity (Appelbaum et al. 1999; Cohen et al. 2004; Lapid et al. 2003). Thus, although intended to safeguard human subjects, concerns about the decisional capacity of persons with mental illness may ironically constitute a form of "benevolent stigmatization" (Synofzik and Clausen 2011) and thereby unnecessarily obstruct research that may lead to effective treatments. Given the limited knowledge about the risks and potential benefits of DBS and the invasive, high-risk nature of DBS research, Dunn and colleagues (Dunn et al. 2011) called for empirical examination of various aspects of the informed consent process to run parallel with the emergence and evolution of DBS clinical trials, and outlined key additional safeguards for protecting participants. Below, we discuss results of this empirical ethics work (Christopher et al. 2012; Fisher et al. 2012; Leykin et al. 2011), highlight some of the outstanding questions and areas for future inquiry, and describe how empirical ethics may similarly guide DBS research for other psychiatric disorders.

DBS for Treatment-Resistant Depression

Early DBS research for psychiatric disorders was focused on TRD because of its high prevalence and devastating effects on patients (Rush et al. 2006; Trivedi et al. 2006). Pilot data suggested that 40–60 % of patients with TRD experienced 50 % or greater reduction in symptom severity when treated with DBS (Bewernick et al. 2010; Lozano et al. 2008; Malone et al. 2009). In addition to its potential for benefit, DBS held other advantages over ablative neurosurgical procedures (also considered for treating refractory psychiatric disease), including stimulation reversibility, adjustability, and revisability. However, although promising, these studies offered limited data on the short-term and no data on the long-term antidepressant efficacy of DBS for TRD.

At the time these pilot data were reported, the known risks associated with DBS were largely extrapolated from research and clinical experience in treating movement disorders. The implantation itself required stereotactic brain surgery in order to place electrodes into the brain and also subcutaneous implantation of a pulse generator/battery pack along the chest wall to power the DBS electrodes. Known risks associated with DBS implantation included intracranial hemorrhage, seizure, infection, hemiparesis, and a small risk of death (Deuschl et al. 2006a, b; Herzog et al. 2003; Synofzik and Schlaepfer 2011). Risks specific to the stimulation site

included increased suicide rate in Parkinson's patients as seen with subthalamic nucleus DBS (Voon et al. 2008), mania in TRD and OCD patients who received anterior internal capsule DBS (Greenberg et al. 2008; Malone et al. 2009), and a number of other mood, cognitive, and functional concerns (Synofzik and Schlaepfer 2011).

There was an expectation that other risks would be identified when larger studies were conducted. Thus, both the known and unknown risks were deemed to be important considerations for any subject with TRD considering enrollment in a DBS trial. Additional factors had to do with maintenance issues associated with DBS implantation, including surgical replacement of implanted battery packs, repair and replacement of dysfunctional DBS components, need for long-term follow-up and monitoring, and inconvenience issues from having to avoid metal detectors, strong magnetic fields (e.g., MRI scanners), and excessive heat (e.g., therapies such as diathermy) (Farris et al. 2008).

The first two multi-site, industry-sponsored clinical trials (Advanced Neuromodulation Systems 2009; MedtronicNeuro 2009) targeted the subcallosal cingulate gyrus (Cg25). The investigators of these trials had to ensure that potential subjects undergo a thorough consent process that included discussion of the known and unknown risks and potential benefits of DBS for TRD, while avoiding exaggeration of potential benefits.

Prospective participants in these two trials were also invited to participate in an add-on study that evaluated the following aspects of their decision-making regarding entering the DBS trial: (1) their capacity to provide informed consent (based on objective measurement of their trial-related understanding, appreciation, reasoning, and expression of a choice); (2) their perceptions of risks and inconveniences, expectations of benefit, and motivations regarding enrollment; and (3) their ability to accurately distinguish the intent and procedures of the trial from ordinary clinical treatment.

Capacity to Consent: Empirical Data

Capacity to provide informed consent to the DBS trial was assessed with the MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR) (Appelbaum and Grisso 2001). The MacCAT-CR requires a semi-structured, open-ended interview that is designed to be adapted to the specific content of any research protocol. Thus, a unique version of the MacCAT-CR was developed for each of the two clinical trials.

The MacCAT-CR is divided into subscales (with higher scores indicating better performance) to rate an individual's ability in four key competence domains: Understanding (i.e., ability to comprehend and recall the disclosed details of the study including its purpose, duration, procedures, risks, and potential benefits), Appreciation (i.e., ability to apply the disclosed information to one's own situation), Reasoning (i.e., ability to critically compare research participation with other treatment options), and Expression of Choice (i.e., clearly indicate and sustain one's own decision regarding enrollment).

Of the individuals who met eligibility criteria for one of the two DBS trials, agreed to participate in the add-on study, and completed the MacCAT-CR assessment, the majority demonstrated sufficient capacity to give informed consent (Fisher et al. 2012). On the Understanding subscale (with a possible range 0–38), all but one subject score 33 or higher and 18 out of 28 subjects scored perfectly. On the Appreciation subscale (with a possible range 0–6), 27 out of 29 subjects scored perfectly, with one subject missing one item and another missing two items. On the Reasoning subscale (with a possible range 0–8), 21 out of 25 subjects scored perfectly, with two subjects missing one item and two other subjects missing two items. All subjects scored perfectly (range 0–2) on Expression of Choice.

It should be noted that this sample represents a very small, generally well-educated and pre-screened group, and that the discussions leading up to an informed consent decision were extensive, often occurring over multiple sessions and lasting for several hours. Nevertheless, the overall high performance of these individuals on the MacCAT-CR undermines the assumption that individuals with TRD lack the requisite decisional capacity to enroll in DBS research. This finding, which is consistent with findings from a number of other studies of decisional capacity among persons with mental illness (Appelbaum et al. 1999; Candilis et al. 2008; Cohen et al. 2004; Grisso and Appelbaum 1995; Jeste et al. 2006; Lapid et al. 2003), suggests that serious mental illness alone should not preclude patients from enrolling in DBS research (or consenting to DBS treatment).

Perceptions of Risks, Expectations of Benefit, and Motivations

Prospective participants in these DBS trials were also asked a series of questions to evaluate their perceptions and ratings of the study's risks (DBS surgery, depression questionnaires, MRI, neuropsychological testing), the potential for personal benefit, the potential for the study to benefit others in the future (by advancing knowledge about treating depression), and their motivation to participate in order to help others (i.e., altruism).

Of the 30 participants that answered these items, nearly all rated the risks associated with completion of depression questionnaires and neuropsychological testing as minimally risky. All participants rated the DBS surgery (implantation) itself as the riskiest part of the study (21 rated it as a "moderate risk" and four as a "minor increase over minimal risk"); no participant rated the surgery as "minimal risk." All but one participant rated the surgery higher than all other study procedures. When comparing how participants rated the risk of each procedure to oneself with people generally, the mean risk ratings did not differ significantly, with the average risk rating being in the "moderate" range (Leykin et al. 2011).

When asked to rate the likelihood of benefiting from the DBS study (from 0 to 5, with a higher score indicating greater likelihood of benefit), on average, participants rated others as being more likely to benefit than themselves (mean of 4.0 (SD = 0.8) vs. 3.21 (SD = 0.6), paired t (27) = −4.53, $p < 0.001$). Participants' ratings of personal risk from the study did correlate with their ratings of study risks to people

in general ($\rho = 0.45$, $p = 0.01$) but did not relate to their ratings of personal benefit ($\rho = 0.09$, $p = 0.63$) (Leykin et al. 2011).

Distinguishing Between Research and Clinical Care

Participants were also asked 8 true/false items intended to assess their ability to accurately distinguish between key ways in which the DBS trial (or research in general) differed from ordinary clinical treatment (i.e., determine whether they had a therapeutic misconception). These included questions about the study's purpose, the likelihood of personally benefiting from the study, and whether the study protocol permitted individualization of treatment for each participant. Although 11 of the 31 participants (35.5 %) gave the correct response for all eight items, the remainder (64.5 %) answered at least one incorrectly. Seven participants (22.6 %) missed two items, and one (3.2 %) missed three items. The items that were most commonly answered incorrectly (by seven participants or 22.5 %) were: "The study is mostly intended to help me" (which is false), and "The research physician cannot add any other medication for depression during the first seven months of the study, even if he or she thinks it would help me" (which is true). Six participants (19 %) also incorrectly indicated that the following item was true: "My own treatment for depression will certainly be improved as a result of this study" (Leykin et al. 2011). Of note, there was no correlation between depression severity and the likelihood of having therapeutic misconception (Fisher et al. 2012).

The prevalence of therapeutic misconception in this group comports with that found among other psychiatric and non-psychiatric research subjects (Appelbaum et al. 2004; Daugherty et al. 1995, 2000; Dunn et al. 2006b; Henderson et al. 2006). It is noteworthy that although these participants do not appear to have unrealistic expectations of personally benefiting from DBS, many nevertheless view the trial as being primarily intended to help them, believe the researchers are able to individualize their treatment, or believe their own depression treatment will certainly improve by being in the study.

Participant's responses to the MacCAT-CR interview were also qualitatively analyzed to examine the influences that guide their decisions regarding enrollment in the DBS trial (Christopher et al. 2012). Not surprisingly, participants cited a wide range of disease- and treatment-related factors. Most commonly were having exhausted existing treatment options (65.4 % of participants), specific ways in which they hoped DBS would improve their lives (84.6 %), inconveniences associated with the study protocol (76.9 %), and specific risks associated with the DBS surgery (53.8 %). Less frequently reported influences included wanting to take initiative in trying to treat their depression (26.9 %), believing that DBS might be more effective because it represented a new and different kind of treatment (30.7 %), hoping their participation would help others or advance science (i.e., altruism) (26.9 %), and the possibility that DBS might be effective (19.2) or not effective (34.6 %). This group's high prevalence of being motivated by a perceived lack of other treatment options and the novelty of the protocol

intervention (i.e., DBS) is consistent with findings on the motives of participants in other early-phase clinical trials (Nurgat et al. 2005; Shannon-Dorcy and Drevdahl 2011). Yet, despite a high number of shared factors, no single consideration seemed to be the only factor in determining one's enrollment decision; indeed, each participant tended to have their own unique set of motives.

DBS for Obsessive-Compulsive Disorder

DBS for obsessive-compulsive disorder (OCD) is also in relatively early development as a treatment for patients who remain symptomatic after multiple treatment trials. OCD is a devastating disorder for these patients, significantly reducing quality of life, curtailing productive and enjoyable daily functioning, and severely constraining the patient's world (Markarian et al. 2010). A significant portion of patients with OCD also suffer from comorbid psychiatric syndromes, most commonly major depression, other anxiety disorders, alcohol use disorders, tics, eating disorders, and other compulsive behaviors (Torres et al. 2006). In addition to the symptoms borne by patients, families suffer alongside patients, with daily work and social functioning constrained by symptoms that can essentially imprison the patient (Cicek et al. 2013).

DBS for OCD presents a set of ethical concerns that overlaps those associated with DBS for depression. There are some differences, however. An important difference is that DBS for OCD has received a humanitarian-device exemption (HDE) from the United States Food and Drug Administration (FDA), meaning that DBS can be used outside of clinical trials for individuals with treatment-refractory OCD. This approval has been criticized by some (Fins et al. 2011a) as premature, given the limited efficacy data available at the time of the approval. Thus, from the perspective of many in the field, despite the HDE approval, DBS remains experimental for OCD. From an ethics perspective, ensuring that patients understand the relatively limited data remains important to ensure that consent is fully informed.

There are no existing empirical studies that have examined decision-making capacity for DBS research in individuals with OCD. In fact, to our knowledge, there are no studies specifically examining decision-making capacity of patients with OCD for any type of clinical research. Nevertheless, concerns about the informed consent process, the desperation of patients with treatment-refractory illness, and the abilities of patients to understand, appreciate, and reason adequately regarding DBS trials parallel those about patients with severe and refractory mood disorders.

Lipsman et al. (2012) enumerated these concerns as follows. First, they noted that patients with treatment-resistant psychiatric conditions, such as depression or OCD, may be vulnerable to "desperation." Unfortunately, however, no operational definition of "desperation" – and no method of measuring this construct – exists in the context of clinical trials. Furthermore, patients with treatment-resistant OCD may be more liable to expect that this novel procedure will benefit them personally, not fully grasping that the intervention is still essentially experimental – even in the presence of an HDE. Lipsman and colleagues express this concern as follows:

> Research, however, into psychiatric DBS, including OCD, is still very much an area under active investigation and as such, a patient's referral to a neurosurgeon typically constitutes the crossing of an invisible border, from treatment to research. This division is rarely explicitly addressed and enhances the vulnerability of these patients by potentially disturbing the consent process, in several ways. (Lipsman et al. 2012, p. 108)

The authors elaborate that patients may view the intervention as an effective treatment when it is essentially still investigational. They also raise the issue of whether referral to a neurosurgeon "without explicit discussion of the therapeutic misconception" (Lipsman et al. 2012) may take advantage of the patient's sense of desperation and in essence confirm to the patient that their condition is desperate. However, no evidence is available to gauge whether these concerns are supported. Until such questions are examined using appropriate, systematic, and thorough methods – particularly in-depth interviews with patients with these disorders (including those considering DBS as well as those not considering DBS), it is nearly impossible to generalize about "desperation," its effects on decision making regarding DBS, or how to address these effects.

Similarly, no research is available regarding levels, predictors, or correlates of decision-making abilities (together comprising capacity to consent) of patients with OCD in the context of research participation decision making. Studies of patients with various psychiatric disorders suggest that the most robust correlates of decision-making abilities, as measured by, e.g., the MacCAT-CR, are cognitive measures, particularly working memory, processing speed, and language abilities (Dunn et al. 2006a, 2007; Palmer et al. 2004; Palmer and Savla 2007; Stroup et al. 2005). Given impairments in emotional processing and cognitive flexibility in patients with OCD, and the detailed nature of the procedures and risks related to DBS, it would be informative for investigators to gather data regarding the decision-making abilities, motivations, risk and benefit perceptions, and potential therapeutic misconception in patients with OCD considering DBS. It would also, in our view, be particularly valuable to examine the actual content and processes of the discussions between the DBS team and the patient, in order to determine whether there are discrepancies between the perspectives and expectations of clinicians and investigators and those of patients.

DBS for Alzheimer's Disease and Other Dementias

The enormous personal, familial, societal, and financial toll of Alzheimer's disease (AD), currently the sixth leading cause of death in the United States (Thies and Bleiler 2011), will only continue to grow, given the projected growth in AD as the population ages (Hebert et al. 2013). Thus far, pharmacologic interventions have been disappointing, with modest effects on the trajectory of the illness. Calls for enhanced research efforts have grown more urgent (The Alzheimer's Study Group 2009). Into this intense environment, a small number of investigators and centers have begun early-phase trials examining the potential for DBS to slow the progression of AD, based, at least initially, on a serendipitous finding of enhanced memory in a single patient implanted for another indication (Hamani et al. 2008).

Ethical issues surrounding DBS for AD (or other disorders characterized by progressive cognitive deterioration, but here we will focus on AD as that is where the trials are now focused) again mirror those for depression and OCD, but with some additional, important issues related to the cognitive impairment that is the hallmark of AD. Specifically, decision-making capacity for research in people with AD is much more likely to be impaired, depending on the stage of disease, and, unlike other psychiatric illnesses, will certainly deteriorate as the illness progresses. Furthermore, consent for research will then need to be a family affair, with surrogate decision makers providing consent with concurrent patient “assent.” These factors raise unique ethical concerns.

First, there is the unresolved question of whether “desperation” experienced by patients and families dealing with this devastating illness may unduly influence people to accept a research intervention that they otherwise would not. Anecdotally, reports in the media suggest that desperation plays a role in research decision making, but again, there are no systematic data to help put such reports into context. There are also no data directly comparing the motivations of patients and families enrolling in DBS research versus other types of AD intervention research, such as pharmacologic or non-pharmacologic trials. There are also no comparative data regarding “desperation” in the context of DBS for dementia versus “desperation” in the context of research for other serious illnesses (e.g., cancer, Parkinson’s, epilepsy, multiple sclerosis, etc.). Such comparisons would be informative, as there may or may not be evidence that the patients and families enrolling in DBS research weigh risks and benefits differently than those enrolling in other types of research. Again, assumptions unsupported by evidence do no favors to the research community nor to the patients and families they ultimately serve. Finally, it remains vital that investigators and others (e.g., those in the media) do not hype the “promise” of this intervention when seeking participants or discussing early results (Racine et al. 2007; Schlaepfer et al. 2010), given the potential propensity of these patients (and other patients with serious disorders) to believe that early “positive” findings indicate efficacy, when in fact much more data are needed.

Second, there is clear evidence from the decision-making capacity literature that people with AD eventually lose the requisite abilities to provide consent to research (Kim 2006; Kim and Caine 2002). The transition from mild to moderate AD appears to be the vulnerable period for loss of decisional capacity for research decision making. Therefore, assessment of decision-making capacity needs to be incorporated into any research protocol that involves people with AD in DBS (or other non-minimal risk) research. Indeed, decisional capacity must be frequently reassessed, since a patient with AD who has decisional capacity to give consent at enrollment may lose that capacity at any subsequent point in the research study. Although periodic re-evaluation of capacity is a generally accepted requirement in any longitudinal study (e.g., “re-consenting” at follow assessments), it is particularly important in the context of DBS for AD, given the potential for adverse effects from DBS that could not reasonably be anticipated at the time initial consent was given. Ideally, then, an initial informed consent procedure should include a discussion of the possibility that the participant will lose decisional capacity

after enrollment. A prospective participant may choose to specify circumstances under which they would want their consent to be revoked or identify a surrogate who would assume responsibility for giving ongoing consent. In one of the initial studies, it is unclear that any capacity assessment was conducted (Laxton et al. 2010), although it was stated that the local ethics review board monitored the informed consent process and "obtained consent independent of the investigators" (Laxton et al. 2010). This is an important point, given that the range of Mini-Mental State Examination scores at the preoperative assessment timepoint was 15–27 (a wide range of cognitive impairment), suggesting that the use of an independent consent procedure was an appropriate safeguard. However, it would also be helpful to know more about the consent procedure, whether any educational intervention strategies were used given the level of cognitive impairment, and whether any standardized, validated capacity assessment method or tool was employed, and whether decisional capacity was reassessed at any point (assuming it was evaluated at the outset). Given the higher risk nature of the intervention, as well as the limited knowledge base about the potential benefits of the intervention for people with AD, the threshold for "adequate" capacity arguably ought to be more stringent than for lower risk protocols. Thus, there remains a need for further attention to the adequacy of the capacity screening and consent procedures for these early-phase studies of DBS for AD.

Third, when patients do lack adequate capacity to provide research consent, investigators must rely on a surrogate decision-maker (most often a family member, such as a spouse or adult child) to provide consent to the protocol (Dunn et al. 2012). In the Laxton study, the investigators do not clearly state how many of the six participants provided their own consent versus how many had surrogates' consent for them (with, presumably, assent provided by the patient). There are numerous ethical dimensions of surrogate consent that need to be considered in any research protocol, but that deserve particular attention in the context of highly invasive and novel interventions. First, ethical guidelines (and some state laws) related to surrogate decision making (Saks et al. 2008) recommend a specific hierarchy of standards for how surrogates should decide on behalf of the patient or subject. First, if there is an advance directive or other evidence of known wishes, the surrogate is supposed to act in accordance with that advance directive. In the absence of an advance directive (as is most frequently the case in research), surrogates are supposed to employ "substituted judgment" – their understanding of what the patient would have wanted if capable. When this is not possible, the surrogate is ideally to decide based on the patient's best interests. A number of studies, however, have shown that, particularly in the context of research participation, surrogates may have difficulty actually using "substituted judgment," given that many people have not expressed premorbid wishes regarding this specific situation (Hare et al. 1992; Hirschman et al. 2006; Seckler et al. 1991; Shalowitz et al. 2006). Instead, surrogates tend to use a combination of best interests and substituted judgment, and describe trying to honor the patient's wishes and values, while trying to maintain the patient's quality of life (Black et al. 2010; Overton et al. 2012). Surrogate decision makers' processes do not neatly reflect the

standards and guidelines that presently exist. For investigators and ethics reviewers, this creates a difficult situation, as they are charged with upholding ethical principles (e.g., "respect for persons" and "beneficence") in a group with significant potential vulnerability. Surrogate consent, it should be noted, does not relieve the investigator of other obligations to uphold ethical duties in research, and may add other duties, such as re-assessment of capacity as well as ongoing monitoring of patient assent to continue participating. In addition, it is not known to what extent, if any, surrogates have a therapeutic misconception with regard to research. This is another important area in need of empirical study.

Conclusion

While deep brain stimulation holds considerable promise in providing relief to a number or debilitating and treatment-refractory psychiatric disorders, the clinical research needed to objectively evaluate this potential is both high risk and, at present, in an early phase. The risks, inconveniences, novelty, and complexities of such research, and the fact that these patients, by definition, have a paucity of alternative options to treat their mental illness, underscore the need for a thorough and thoughtful approach to engage them in informed consent discussions to participate in clinical DBS research. The existing empirical work examining the informed consent process and decision making for enrollment in DBS research, which focuses exclusively on DBS for TRD, suggests that, under the current informed consent process, the majority or eligible participants have adequate decision-making capacity and understanding of the risks associated with DBS. Many participants nevertheless seem to have misconceptions about the nature of the clinical trial, confusing aspects of the DBS study with routine clinical treatment. Although not specific to this research population, the high prevalence of this phenomenon is worrisome, given the extensive nature of the informed consent discussions required for DBS trial consideration. On closer inspection, participants also describe idiosyncratic and varied factors that guide their enrollment decisions. Taken together, these findings suggest a more personalized approach to informed consent to DBS trials may be needed, beginning with a discussion that seeks to elicit prospective participants' motives, fears, and hopes for considering a DBS trial, and that carefully probes and corrects for therapeutic misconceptions in addition to ensuring their requisite understanding of the trial's procedures, risks, and potential benefits.

In the context of DBS for OCD or AD, there are no empirical data specifically examining capacity to consent to DBS research; therefore, any conclusions are necessarily speculative. However, concerns about patients' and families' "desperation" for treatment surround DBS for these disorders as well. Moreover, for disorders characterized by cognitive impairment, such as AD, additional safeguards related to capacity assessment and surrogate consent should be considered. Attention to addressing participants' and families' personal concerns, as well as trying to understand motivations for participating, should be the focus of future empirical studies in DBS research ethics.

As potential disease targets for DBS expand – e.g., to addictions or eating disorders, as well as potentially to less severe forms of mood and cognitive disorders – we anticipate that there will be intensified focus on these and other ethical issues that relate to the nature of the intervention and the types of disorders that are the subject of study. As with all research involving human subjects, continual vigilance of those entrusted with the scientific and ethical conduct and oversight of research will be required. Further conceptual and empirical work will also be needed to continue to inform discussion and debate regarding the ethical application of these novel therapeutics.

Cross-References

▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
▶ Ethical Issues in the Treatment of Addiction
▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
▶ Ethics in Neurosurgery
▶ Ethics in Psychiatry
▶ Ethics of Functional Neurosurgery
▶ Ethics of Pharmacological Mood Enhancement
▶ Experimentation in Cognitive Neuroscience and Cognitive Neurobiology
▶ Informed Consent and the History of Modern Neurosurgery
▶ Relationship of Benefits to Risks in Psychiatric Research Interventions

References

Advanced Neuromodulation Systems, Inc. (2009). *Broaden clinical study*.

Appelbaum, P. S., & Grisso, T. (2001). *MacCAT-CR: MacArthur Competence Assessment Tool for Clinical Research*. Sarasota, FL: Professional Resource Press.

Appelbaum, P. S., Grisso, T., Frank, E., O'Donnell, S., & Kupfer, D. J. (1999). Competence of depressed patients for consent to research. *American Journal of Psychiatry, 156*(9), 1380–1384.

Appelbaum, P. S., Lidz, C. W., & Grisso, T. (2004). Therapeutic misconception in clinical research: Frequency and risk factors. *IRB, 26*(2), 1–8.

Bell, E., Mathieu, G., & Racine, E. (2009). Preparing the ethical future of deep brain stimulation. *Surgical Neurology, 72*(6), 577–586. doi:10.1016/j.surneu.2009.03.029. S0090-3019(09)00285-7 [pii].

Bewernick, B. H., Hurlemann, R., Matusch, A., Kayser, S., Grubert, C., Hadrysiewicz, B., Schlaepfer, T. E. (2010). Nucleus accumbens deep brain stimulation decreases ratings of depression and anxiety in treatment-resistant depression. *Biological Psychiatry, 67*(2), 110–116. doi:10.1016/j.biopsych.2009.09.013, S0006-3223(09)01094-4 [pii].

Black, B. S., Rabins, P. V., Sugarman, J., & Karlawish, J. H. (2010). Seeking assent and respecting dissent in dementia research. *American Journal of Geriatric Psychiatry, 18*(1), 77–85. doi:10.1097/JGP.0b013e3181bd1de2.

Candilis, P. J., Fletcher, K. E., Geppert, C. M., Lidz, C. W., & Appelbaum, P. S. (2008). A direct comparison of research decision-making capacity: Schizophrenia/schizoaffective, medically ill, and non-ill subjects. *Schizophrenia Research, 99*(1–3), 350–358. doi:10.1016/j.schres.2007.11.022. S0920-9964(07)00534-8 [pii].

Christopher, P. P., Leykin, Y., Appelbaum, P. S., Holtzheimer, P. E., 3rd, Mayberg, H. S., & Dunn, L. B. (2012). Enrolling in deep brain stimulation research for depression: Influences on potential subjects' decision making. *Depression and Anxiety, 29*, 139–146. doi:10.1002/da.20916.

Cicek, E., Cicek, I. E., Kayhan, F., Uguz, F., & Kaya, N. (2013). Quality of life, family burden and associated factors in relatives with obsessive-compulsive disorder. *General Hospital Psychiatry, 35*(3), 253–8. doi:10.1016/j.genhosppsych.2013.01.004. S0163-8343(13)00009-1 [pii].

Cohen, B. J., McGarvey, E. L., Pinkerton, R. C., & Kryzhanivska, L. (2004). Willingness and competence of depressed and schizophrenic inpatients to consent to research. *Journal of the American Academy of Psychiatry and the Law, 32*(2), 134–143.

Daugherty, C., Ratain, M. J., Grochowski, E., Stocking, C., Kodish, E., Mick, R., & Siegler, M. (1995). Perceptions of cancer patients and their physicians involved in phase i trials. *Journal of Clinical Oncology, 13*(5), 1062–1072.

Daugherty, C. K., Banik, D. M., Janish, L., & Ratain, M. J. (2000). Quantitative analysis of ethical issues in phase i trials: A survey interview of 144 advanced cancer patients. *IRB, 22*(3), 6–14.

Deuschl, G., Herzog, J., Kleiner-Fisman, G., Kubu, C., Lozano, A. M., Lyons, K. E., Voon, V. (2006a). Deep brain stimulation: Postoperative issues. *Movement Disorders, 21*(Suppl. 14), S219–S237.

Deuschl, G., Schade-Brittinger, C., Krack, P., Volkmann, J., Schafer, H., Botzel, K., Voges, J. (2006b). A randomized trial of deep-brain stimulation for Parkinson's disease. *New England Journal of Medicine, 355*(9), 896–908.

Dunn, L. B., Candilis, P. J., & Roberts, L. W. (2006a). Emerging empirical evidence on the ethics of schizophrenia research. *Schizophrenia Bulletin, 32*(1), 47–68. doi:10.1093/schbul/sbj012.

Dunn, L. B., Palmer, B. W., Keehan, M., Jeste, D. V., & Appelbaum, P. S. (2006b). Assessment of therapeutic misconception in older schizophrenia patients with a brief instrument. *American Journal of Psychiatry, 163*(3), 500–506. doi:10.1176/appi.ajp.163.3.500.

Dunn, L. B., Palmer, B. W., Appelbaum, P. S., Saks, E. R., Aarons, G. A., & Jeste, D. V. (2007). Prevalence and correlates of adequate performance on a measure of abilities related to decisional capacity: Differences among three standards for the MacCAT-CR in patients with schizophrenia. *Schizophrenia Research, 89*(1–3), 110–118. doi:10.1016/j.schres.2006.08.005.

Dunn, L. B., Holtzheimer, P. E., 3rd, Hoop, J. G., Mayberg, H., Roberts, L. W., & Appelbaum, P.S. (2011). Ethical issues in deep brain stimulation research for treatment-resistant depression: Focus on risk and consent. *AJOB Neuroscience, 2*, 29–36.

Dunn, L. B., Fisher, S. R., Hantke, M., Appelbaum, P. S., Dohan, D., Young, J. P., & Roberts, L. W. (2013). "Thinking about it for somebody else": Alzheimer's disease research and proxy decision makers' translation of ethical principles into practice. *American Journal of Geriatric Psychiatry 21,* 337–345. doi:10.1097/JGP.0b013e31824362ca

Elliott, C. (1997). Caring about risks: Are severely depressed patients competent to consent to research? *Archives of General Psychiatry, 54*, 113–116.

Farris, S., Vitek, J., & Giroux, M. L. (2008). Deep brain stimulation hardware complications: The role of electrode impedance and current measurements. *Movement Disorders, 23*(5), 755–760. doi:10.1002/mds.21936.

Fins, J. J., Mayberg, H. S., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., Schlaepfer, T. E. (2011a). Misuse of the FDA'S humanitarian device exemption in deep brain stimulation for obsessive-compulsive disorder. *Health Affairs, 30*(2), 302–311. doi:10.1377/hlthaff.2010.0157, 30/2/302 [pii]

Fins, J. J., Schlaepfer, T. E., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., Mayberg, H. S. (2011b). Ethical guidance for the management of conflicts of interest for researchers, engineers and clinicians engaged in the development of therapeutic deep brain stimulation. *Journal of Neural Engineering*, *8*(3), 033001. doi:10.1088/1741-2560/8/3/033001, S1741-2560(11) 82782-5 [pii].

Fisher, C. E., Dunn, L. B., Christopher, P. P., Holtzheimer, P. E., Leykin, Y., Mayberg, H. S., Appelbaum, P. S. (2012). The ethics of research on deep brain stimulation for depression: Decisional capacity and therapeutic misconception. *Annals of the New York Academy of Sciences, 1265*, 69–79. doi:10.1111/j.1749-6632.2012.06596.x.

Flaherty, A. W., Williams, Z. M., Amirnovin, R., Kasper, E., Rauch, S. L., Cosgrove, G. R., & Eskandar, E. N. (2005). Deep brain stimulation of the anterior internal capsule for the treatment of Tourette syndrome: Technical case report. *Neurosurgery, 57*(Suppl. 4), E403, 00006123-200510004-00029 [pii]; discussion E403.

Goodman, W. K., Foote, K. D., Greenberg, B. D., Ricciuti, N., Bauer, R., Ward, H., Okun, M. S. (2010). Deep brain stimulation for intractable obsessive compulsive disorder: Pilot study using a blinded, staggered-onset design. *Biological Psychiatry*, *67*(6), 535–542. doi:10.1016/j.biopsych.2009.11.028, S0006-3223(09)01426-7 [pii].

Greenberg, B. D., Malone, D. A., Friehs, G. M., Rezai, A. R., Kubu, C. S., Malloy, P. F., Rasmussen, S. A. (2006). Three-year outcomes in deep brain stimulation for highly resistant obsessive-compulsive disorder. *Neuropsychopharmacology, 31*(11), 2384–2393.

Greenberg, B. D., Gabriels, L. A., Malone, D. A., Jr., Rezai, A. R., Friehs, G. M., Okun, M. S., Nuttin, B. J. (2008). Deep brain stimulation of the ventral internal capsule/ventral striatum for obsessive-compulsive disorder: Worldwide experience. *Molecular Psychiatry, 15*(1):64–79. doi:10.1038/mp.2008.55, mp200855 [pii].

Grisso, T., & Appelbaum, P. S. (1995). The MacArthur treatment competence study. III: Abilities of patients to consent to psychiatric and medical treatments. *Law and Human Behavior, 19*(2), 149–174.

Hamani, C., McAndrews, M. P., Cohn, M., Oh, M., Zumsteg, D., Shapiro, C. M., Lozano, A. M. (2008). Memory enhancement induced by hypothalamic/fornix deep brain stimulation. *Annals of Neurology, 63*(1), 119–123. doi:10.1002/ana.21295

Hare, J., Pratt, C., & Nelson, C. (1992). Agreement between patients and their self-selected surrogates on difficult medical decisions. *Archives of Internal Medicine, 152*(5), 1049–1054.

Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 Census. *Neurology, 80*(19), 1778–83. doi:10.1212/WNL.0b013e31828726f5. WNL.0b013e31828726f5 [pii].

Henderson, G. E., Easter, M. M., Zimmer, C., King, N. M., Davis, A. M., Rothschild, B. B., Nelson, D. K. (2006). Therapeutic misconception in early phase gene transfer trials. *Social Science and Medicine, 62*(1), 239–253.

Herzog, J., Volkmann, J., Krack, P., Kopper, F., Potter, M., Lorenz, D., Deuschl, G. (2003). Two-year follow-up of subthalamic deep brain stimulation in Parkinson's disease. *Movement Disorders, 18*(11), 1332–1337

Hirschman, K. B., Kapo, J. M., & Karlawish, J. H. (2006). Why doesn't a family member of a person with advanced dementia use a substituted judgment when making a decision for that person? *American Journal of Geriatric Psychiatry, 14*(8), 659–667.

Holtzheimer, P. E., & Mayberg, H. S. (2011). Deep brain stimulation for psychiatric disorders. *Annual Review of Neuroscience, 34*, 289–307. doi:10.1146/annurev-neuro-061010-113638.

Holtzheimer, P. E., Kelley, M. E., Gross, R. E., Filkowski, M. M., Garlow, S. J., Barrocas, A., Mayberg, H. S. (2012). Subcallosal cingulate deep brain stimulation for treatment-resistant unipolar and bipolar depression. *Archives of General Psychiatry, 69*(2), 150–158. doi:10.1001/archgenpsychiatry.2011.1456.

Jeste, D. V., Depp, C. A., & Palmer, B. W. (2006). Magnitude of impairment in decisional capacity in people with schizophrenia compared to normal subjects: An overview. *Schizophrenia Bulletin, 32*(1), 121–128.

Kennedy, S. H., Giacobbe, P., Rizvi, S. J., Placenza, F. M., Nishikawa, Y., Mayberg, H. S., & Lozano, A. M. (2011). Deep brain stimulation for treatment-resistant depression: Follow-up after 3 to 6 years. *American Journal of Psychiatry, 168*(5), 502–510. doi:10.1176/appi.ajp.2010.10081187. appi.ajp.2010.10081187 [pii].

Kim, S. Y. (2006). When does decisional impairment become decisional incompetence? Ethical and methodological issues in capacity research in schizophrenia. *Schizophrenia Bulletin, 32*(1), 92–97.

Kim, S. Y., & Caine, E. D. (2002). Utility and limits of the mini mental state examination in evaluating consent capacity in Alzheimer's disease. *Psychiatric Services, 53*(10), 1322–1324.

Lapid, M. I., Rummans, T. A., Poole, K. L., Pankratz, V. S., Maurer, M. S., Rasmussen, K. G., Appelbaum, P. S. (2003). Decisional capacity of severely depressed patients requiring electroconvulsive therapy. *The Journal of ECT, 19*(2), 67–72.

Laxton, A. W., & Lozano, A. M. (2012). Deep brain stimulation for the treatment of Alzheimer disease and dementias. *World Neurosurgery*. doi:10.1016/j.wneu.2012.06.028.

Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Zumsteg, D., Wennberg, R., Keren, R., Lozano, A. M. (2010). A phase i trial of deep brain stimulation of memory circuits in Alzheimer's disease. *Annals of Neurology, 68*(4), 521–534. doi:10.1002/ana.22089.

Leykin, Y., Christopher, P. P., Holtzheimer, P. E., Appelbaum, P. S., Mayberg, H. S., Lisanby, S. H., & Dunn, L. B. (2011). Participants' perceptions of deep brain stimulation research for treatment-resistant depression: Risks, benefits, and therapeutic misconception. *AJOB Primary Research, 2*, 33–41.

Lipsman, N., Giacobbe, P., Bernstein, M., & Lozano, A. M. (2012). Informed consent for clinical trials of deep brain stimulation in psychiatric disease: Challenges and implications for trial design. *Journal of Medical Ethics, 38*, 107–111. doi:10.1136/jme.2010.042002.

Lipsman, N., Woodside, D. B., Giacobbe, P., Hamani, C., Carter, J. C., Norwood, S. J., Lozano, A. M. (2013). Subcallosal cingulate deep brain stimulation for treatment-refractory anorexia nervosa: A phase 1 pilot trial. *Lancet*, *381*(9875):1361–1370. doi:10.1016/S0140-6736(12)62188-6, S0140-6736(12)62188-6 [pii].

Lozano, A. M., Mayberg, H. S., Giacobbe, P., Hamani, C., Craddock, R. C., & Kennedy, S. H. (2008). Subcallosal cingulate gyrus deep brain stimulation for treatment-resistant depression. *Biological Psychiatry, 64*(6), 461–467.

Lozano, A. M., Giacobbe, P., Hamani, C., Rizvi, S. J., Kennedy, S. H., Kolivakis, T. T., Mayberg, H. S. (2012). A multicenter pilot study of subcallosal cingulate area deep brain stimulation for treatment-resistant depression. *Journal of Neurosurgery, 116*(2), 315–322. doi:10.3171/2011.10.JNS102122.

Lyons, M. K. (2011). Deep brain stimulation: Current and future clinical applications. *Mayo Clinic Proceedings, 86*(7), 662–672. doi:10.4065/mcp.2011.0045.

Malone, D. A., Jr., Dougherty, D. D., Rezai, A. R., Carpenter, L. L., Friehs, G. M., Eskandar, E. N., Greenberg, B. D. (2009). Deep brain stimulation of the ventral capsule/ventral striatum for treatment-resistant depression. *Biological Psychiatry, 65*(4), 267–275.

Markarian, Y., Larson, M. J., Aldea, M. A., Baldwin, S. A., Good, D., Berkeljon, A., McKay, D. (2010). Multiple pathways to functional impairment in obsessive-compulsive disorder. *Clinical Psychology Review*, *30*(1), 78–88. doi:10.1016/j.cpr.2009.09.005, S0272-7358(09)00136-6 [pii].

Mayberg, H. S., Lozano, A. M., Voon, V., McNeely, H. E., Seminowicz, D., Hamani, C., Kennedy, S. H. (2005). Deep brain stimulation for treatment-resistant depression. *Neuron, 45*(5), 651–660.

MedtronicNeuro. (2009). Reclaim deep brain stimulation clinical study for treatment-resistant depression. Clinicaltrials.Gov identifier. Nct00837486.

Michels, R. (2004). Research on persons with impaired decision making and the public trust. *The American Journal of Psychiatry, 161*(5), 777–779.

National Bioethics Advisory, Commission. (1998). Research involving persons with mental disorders that may affect decision-making capacity. Rockville, MD.: National Bioethics Advisory Commission.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *Report and Recommendations: Research Involving Those Institutionalized*

As Mentally Infirm. Washington, D.C.: U.S. Government Printing Office, 1978. DHEW Publication No. (OS) 78–0006.

Nurgat, Z. A., Craig, W., Campbell, N. C., Bissett, J. D., Cassidy, J., & Nicolson, M. C. (2005). Patient motivations surrounding participation in phase i and phase ii clinical trials of cancer chemotherapy. *British Journal of Cancer, 92*(6), 1001–1005. doi:10.1038/sj.bjc.6602423. 6602423 [pii].

Overton, E., Appelbaum, P. S., Fisher, S. R., Dohan, D., Roberts, L. W., & Dunn, L. B. (2012). Alternative decision-makers' perspectives on assent and dissent for dementia research. *American Journal of Geriatric Psychiatry, 21*(4), 346–354. doi:10.1097/JGP.0b013e318265767e.

Palmer, B. W., & Savla, G. N. (2007). The association of specific neuropsychological deficits with capacity to consent to research or treatment. *Journal of International Neuropsychological Society, 13*(6), 1047–1059.

Palmer, B. W., Dunn, L. B., Appelbaum, P. S., & Jeste, D. V. (2004). Correlates of treatment-related decision-making capacity among middle-aged and older patients with schizophrenia. *Archives of General Psychiatry, 61*(3), 230–236. doi:10.1001/archpsyc.61.3.230.

Rabins, P., Appleby, B. S., Brandt, J., DeLong, M. R., Dunn, L. B., Gabriels, L., Mathews, D. J. (2009). Scientific and ethical issues related to deep brain stimulation for disorders of mood, behavior, and thought. *Archives of General Psychiatry, 66*(9), 931–937. doi:10.1001/archgenpsychiatry.2009.113

Racine, E., Waldman, S., Palmour, N., Risse, D., & Illes, J. (2007). "Currents of hope": Neurostimulation techniques in U.S. and U.K. print media. *Cambridge Quarterly of Healthcare Ethics, 16*(3), 312–316.

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry, 163*(11), 1905–1917.

Saks, E. R., Dunn, L. B., Wimer, J., Gonzales, M., & Kim, S. (2008). Proxy consent to research: The legal landscape. *Yale Journal of Health Policy, Law, and Ethics, 8*(1), 37–92.

Schlaepfer, T. E., & Fins, J. J. (2010). Deep brain stimulation and the neuroethics of responsible publishing: When one is not enough. *JAMA, 303*(8), 775–776. doi:10.1001/jama.2010.140. 303/8/775 [pii].

Schlaepfer, T. E., Lisanby, S. H., & Pallanti, S. (2010). Separating hope from hype: Some ethical implications of the development of deep brain stimulation in psychiatric research and treatment. *CNS Spectrums, 15*(5), 285–287.

Schuepbach, W. M., Rau, J., Knudsen, K., Volkmann, J., Krack, P., Timmermann, L., Deuschl, G. (2013). Neurostimulation for Parkinson's disease with early motor complications. *New England Journal of Medicine, 368*(7), 610–622. doi:10.1056/NEJMoa1205158.

Seckler, A. B., Meier, D. E., Mulvihill, M., & Paris, B. E. (1991). Substituted judgment: How accurate are proxy predictions? *Annals of Internal Medicine, 115*(2), 92–98.

Shalowitz, D. I., Garrett-Mayer, E., & Wendler, D. (2006). The accuracy of surrogate decision makers: A systematic review. *Archives of Internal Medicine, 166*(5), 493–497.

Shannon-Dorcy, K., & Drevdahl, D. J. (2011). "I had already made up my mind": Patients and caregivers' perspectives on making the decision to participate in research at a us cancer referral center. *Cancer Nursing, 34*(6), 428–433. doi:10.1097/NCC.0b013e318207cb03.

Siegfried, J., Lazorthes, Y., & Sedan, R. (1980). Indications and ethical considerations of deep brain stimulation. *Acta Neurochirurgica Supplement, 30*, 269–274.

Stroup, S., Appelbaum, P., Swartz, M., Patel, M., Davis, S., Jeste, D., Lieberman, J. (2005). Decision-making capacity for research participation among individuals in the CATIE schizophrenia trial. *Schizophrenia Research, 80*(1), 1–8.

Synofzik, M., & Clausen, J. (2011). The ethical differences between psychiatric and neurologic dbs: Smaller than we think? *AJOB Neuroscience, 2*(1).

Synofzik, M., & Schlaepfer, T. E. (2011). Electrodes in the brain–ethical criteria for research and treatment with deep brain stimulation for neuropsychiatric disorders. *Brain Stimulation, 4*(1), 7–16. doi:10.1016/j.brs.2010.03.002.

The Alzheimer's Study Group. (2009). *A national Alzheimer's strategic plan: The report of the Alzheimer's Study Group*. Washington, DC: Alzheimer's Association.

Thies, W., & Bleiler, L. (2011). 2011 Alzheimer's disease facts and figures. *Alzheimer's Dement, 7*(2), 208–244. doi:10.1016/j.jalz.2011.02.004.

Torres, A. R., Prince, M. J., Bebbington, P. E., Bhugra, D., Brugha, T. S., Farrell, M., Singleton, N. (2006). Obsessive-compulsive disorder: Prevalence, comorbidity, impact, and help-seeking in the British National Psychiatric Morbidity Survey of 2000. *American Journal of Psychiatry*, *163*(11), 1978–1985. doi:10.1176/appi.ajp.163.11.1978, 163/11/1978 [pii].

Trivedi, M. H., Fava, M., Wisniewski, S. R., Thase, M. E., Quitkin, F., Warden, D., Rush, A. J. (2006). Medication augmentation after the failure of SSRIs for depression. *The New England Journal of Medicine*, *354*(12), 1243–1252. doi:10.1056/NEJMoa052964, 354/12/1243 [pii].

Voon, V., Krack, P., Lang, A. E., Lozano, A. M., Dujardin, K., Schupbach, M., Moro, E. (2008). A multicentre study on suicide outcomes following subthalamic stimulation for Parkinson's disease. *Brain*, *131*(Pt. 10), 2720–2728. doi:10.1093/brain/awn214, awn214 [pii].

Weaver, F. M., Follett, K. A., Stern, M., Luo, P., Harris, C. L., Hur, K., Reda, D. J. (2012). Randomized trial of deep brain stimulation for Parkinson disease: Thirty-six-month outcomes. *Neurology*, *79*(1), 55–65. doi:10.1212/WNL.0b013e31825dcdc1, WNL.0b013e31825dcdc1 [pii].

Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine

37

Joseph J. Fins

Contents

Abstract

In this chapter, the specific challenges of device regulation are addressed with the suggestion that this analysis might be optimized by analogy to personalized medicine, now ascendant in molecular therapeutics. The challenge of device versus pharmaceutical regulation is made through the example of deep brain stimulation (DBS). Unlike drugs, devices are few in number, expensive to develop and administer, and require discrete medical-surgical interdisciplinary expertise. They are personalized in their scope, are therapeutic and investigative tools, and uniquely face barriers related to intellectual property exchange and conflicts of interest. The recent controversy over FDA's Humanitarian Device Exemption (HDE) of DBS in the obsessive-compulsive disorder is offered as evidence of regulatory insufficiency and the failure to distinguish research from therapy due to the therapeutic misconception. An alternative fiscally and methodologically viable research pathway is advanced for quality DBS research.

J.J. Fins
Division of Medical Ethics, New York Presbyterian–Weill Cornell Medical Center, Weill Medical College of Cornell University, New York, NY, USA

Consortium for the Advanced Study of Brain Injury (CASBI), Weill Cornell Medical College & Rockefeller University, New York, NY, USA
e-mail: jjfins@med.cornell.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_33,

The conclusion is on a hopeful note, with the speculation that the advent of personalized medicine in other investigative realms may provide innovative solutions to regulation which promote scientific discovery and meet patient-centered needs.

A Pharma Provenance

Although device regulation is within the purview of the US Food and Drug Administration, this responsibility is notably not in the agency's title. This is more than a semantic omission. It is reflective of the agency's history and far greater role in the regulation of pharmaceuticals (Carpenter 2010), a fact which has consequences for the oversight of devices which follow a regulatory pathway similar to that of drugs, even though there are fundamental differences between drugs and devices.

One of the salient differences between these two sectors is their relative size. Big pharma dwarfs the device industry in economic impact and the number of companies operating in the space. As of this writing in late November 2012, Medtronic has the largest market capitalization of any pure device manufacturer with a value of $43.61 billion. To put this figure into perspective, there are seven pharmaceutical houses with market caps over $100 billion each (Reuters 2012). This figure includes Johnson & Johnson which makes drugs, devices, and consumer products and had $25.8 billion sales in its medical device and diagnostic segment in 2011. This is coupled with $24.4 and $14.9 billion in sales from its pharmaceutical and consumer markets (Johnson and Johnson 2011).

Notwithstanding Johnson & Johnson and its presence in multiple marketplaces, the difference in financial scope between the drug and device sector is quite significant. For example, the combined value of the *top 22* device manufacturers with market caps over $1 billion is *less* than the value of Pfizer, a pharmaceutical company valued at $179.2 billion, or Johnson & Johnson with its more blended portfolio and a market cap of $192.2 billion (Reuters 2012).

The relative size of the sectors influences oversight with the importance of drug regulation quantitatively dominating the regulation of devices and qualitatively informing models of regulation. As we shall see, one of the challenges confronting those who would regulate medical devices is an over reliance on approaches and methods borrowed or derived from the regulation of the pharmaceutical industry. This pharma provenance is unfortunate because there are salient differences between drugs and devices in their development, deployment, prevalence, and markets. In this chapter, I will explore the specific challenges of device regulation (Fins et al. 2011d) through the prism of deep brain stimulators employed in the treatment and study of neuropsychiatric conditions. I will primarily focus on the experience in the United States appreciating that there are salient differences in other jurisdictions.

Although some of the enumerated challenges are unique to deep brain stimulation, much of the critique is generic and could be applied to other areas of device regulation where methods derived from pharma are applied without a thorough appreciation that devices are different from drugs and that the regulation of one sector cannot be utilized on the other without potentially adverse consequences. It is concluded that the advent of personalized medicine may help bridge the regulatory gulf separating devices and drugs.

Drugs, Devices, and Difference

If we are to assert that devices and drugs each require a distinct brand of regulation, it is important to delineate the differences between these therapeutic modalities. Let us start with therapeutic impact and how efficacy relates to cost and prevalence. In general, drugs cost less and have a wider population impact than devices which have a higher start-up capital versus operational cost which may limit dissemination, even though cost-effectiveness analysis reveals that DBS in Parkinson's disease is competitive with pharmacotherapy in a European context (Valldeoriola et al. 2007).

From a revenue standpoint, the low unit cost and broad dissemination allow a drug to be considered a success if it has a small effect on a large number of people. While drugs like antibiotics have a low cost and have dramatic and immediate effects, most of the pharma marketplace is now composed of modern-day agents that address the threats of chronic disease illness. These are not the "miracle cures" or "silver bullets" of yore, but drugs like cholesterol-lowering statins or antihypertensive agents, which generate revenue on a daily basis, one pill at a time. Expectations for these drugs are tempered because their action is not the cure of disease but rather its prevention or amelioration, pushing back the threat of chronic diseases like heart disease or hypertension.

Expectations and revenue streams are quite different for devices which have a high unit cost. DBS electrodes and insertion can cost over $50,000. Given this, to be considered a success and market worthy, their expected impact has to be larger for each individual and more immediately apparent. A quick and significant therapeutic payoff is the marker of success because far fewer devices will be implanted than drugs and because the cost of the device – and its surgical implantation – exceeds the routine cost of drugs. In addition, because all the costs are front loaded, there is the expectation that the treatment effect will come quickly and be more surgical than medical, that is, large and immediate versus incremental and delayed.

Put another way, most drugs are rather utilitarian having a small effect on a large number of people over time. In the aggregate they positively affect the public health by reducing heart disease or stroke, as in the case of lipid-lowering agents or blood pressure drugs. (Conversely, over time, their societal misuse can have an adverse effect in the aggregate by prompting antimicrobial drug resistance with an ecosystem (Fins 1995)).

Devices are more personalized in their scope and impact. They are deployed less frequently. Although there are cogent medical and ethical arguments for their early deployment (Synofzik and Schlaepfer 2011; Schuepbach et al. 2013), historically they have generally been a treatment of last resort, to borrow a phrase from the late Jack Pressman who wrote a seminal volume on the early history of psychosurgery (Pressman 1998). In a twist on the usual way we think about genomic-/molecular-based "personalized medicine," deep brain stimulators are quite personalized, tailored to both the biology of a specific neuropsychiatric malady and the *particular* anatomy of a specific patient. Hence, a very major difference between drugs and devices is their delivery systems. Save for specialized infusions and personalized medicine, mass-produced drugs come from the pharmacy and are taken with a glass of water. They are standardized and come with expected bioavailabilities.

Devices, in contrast, require the intervention of a highly skilled surgeon as well as the expertise of an interdisciplinary team to program the device (Fins et al. 2006), *personalizing* both the assessment of the patient, the insertion of the device, and its manipulation. While medical/pharmacological interventions can require similar skill, the potential variance of effect based on surgical placement and device programming makes this a more personal engagement than pharmacology alone.

Moreover, the presence of the surgical intermediary, essential to device delivery, also is an important regulatory issue. It puts the surgeon in a central role both for his or her technical expertise and a purveyor of goods and as a purchaser. In the context of devices, the surgeon who implants the device becomes a consumer, if not in actual practice when his or her hospital purchases devices, then due to the influence that can be wielded to direct institutional purchasing.

The intimacy of this market, surgeons and vendors, also creates the opportunity for conflicts of interest (Fins 2007), marketing outreach, and research support that, at least in the case of spinal implants, has drawn investigations and Congressional scrutiny and serious concerns about impropriety and conflicts of interest (Armstrong 2008; Abelson 2007). As yet, this scandal has not breached the tentorium and the realm of DBS, but the context for its entry is there and worthy of concern and heightened oversight (Fins and Schiff 2010).

The nature of the small market also creates the opportunity for monopolistic practices and favoritism. The small number of device manufacturers can limit open access to devices for clinical trials because the industry favors those surgeons who are high volume users of their devices. This limits access to other investigators, either by withholding products or by denying "rights of reference," information about a device that an investigator would need to make an application to the FDA for study approval. Without this technical information, there is no way to submit, much less be approved, for a device trial (Synofzik and Schlaepfer 2011; Fins 2007, 2010).

This is more than a semantic point because these devices differ from drugs as a tool of inquiry. It is asserted that deep brain stimulators, unlike drugs, are *probative* as a means to discovery as they are therapeutic agents (Fins 2012). At this early juncture in their development, and our understanding of the therapeutic potentiality of the circuitry of the brain, all DBS work, approved as in the treatment of refractory

Parkinson's disease or not (Bronstein et al. 2011), remain ripe for the cultivation of scientific knowledge. This is quite distinct from the use of an "approved" drug whose utilization represents the culmination of a step-by-step sequence through an experimental process resulting in an agent that becomes a vetted therapy.

The sequence of drug discovery, in contrast to devices, is also laid out and its steps are understood. How we come to understand the circuitry of the brain and its therapeutic potential and its role in pathogenesis is far less stereotyped. In drug development the approach has become formulaic. Phase I trials demonstrate safety. Phase II studies demonstrate benefit and Phase III protocols show that the new agent is as good or better than the standard of care either in their therapeutic effects or an enhanced side-effect profile. The more recent Phase IV seeks post-marketing data for evidence of late-breaking side effects or toxicities, not apparent in earlier stages of assessment.

Although some "best practice" device trials have utilized the "Phase I–IV" demarcations (Laxton et al. 2010), this is far less universal than its usage with respect to the evolution of drug trials. And when such designations are used, they are simply derivative from the experience with drugs and pharma. Devices do not have their own schema, notwithstanding the occasional invocation of the phase milestones to studies.

Indeed, no such generally agreed-upon sequence exists for devices, making for confusion on how to move along a bright idea from bench to bedside and clearly demarcating the assessment of safety and efficacy. And the reason no clear sequence exists is because the use of DBS remains highly experimental and contingent upon the interplay of device (whose actions are predictable) and soma (which is variable dependent on anatomic locale, disease state, and programming of stimulation parameters) (Fins 2004). More fundamentally, the use of DBS remains fundamental to garnering information at the frontiers of human knowledge. The use of these devices is not yet ever exclusively therapeutic but always in part investigational elucidating the circuitry of the human brain and our ability to harness it or redirect it in the pursuit of health and well-being (Fins 2012). To view DBS, or much less regulate it, solely as a therapeutic device is to rob the work of its investigational potential, a potential that does not reside in most drugs approved by FDA.

And this leads to the final and perhaps most critical dissimilarity between the regulation of drugs and DBS, the role that intellectual property transfer plays in sustaining and promoting innovation. In the United States, intellectual property (IP) exchange between industry and the academy is governed by the Bayh-Dole Act of 1980, legislation that allows the rights to federally sponsored research which results in IP to be transferred from the US Government to the University where the pioneering work was done (Bayh-Dole Act 1980). The premise behind the law was that IP exchange to an interested and incentivized party (the University with its faculty investigator) would expedite negotiations with industry and accelerate the process of moving innovation from bench to bedside. And universities holding IP would be in a position to negotiate with industry because drug development could not go forward without pharma

owning the IP. In the aggregate this exchange of IP would bring new drugs most quickly to those in need.

In theory, and indeed in some quarters in practice, this legislation has fostered innovation. But in the context of DBS devices, it has, in my view, hindered progress by commodifying IP and giving it a monetary value too early in the course of development (Fins 2010). This is problematic because this commodification burdens the investigator with a putative conflict of interest that can preclude, under some guidelines, frontline research, research that they are most likely the best positioned to do. I suggested that IP transfer be delayed until after proof of principle in order to allow investigators most familiar with the work to get their studies through this pivotal early stage without placing a valuation on their ideas (Fins 2010).

This is suggested, in part, because of conflict of interest policies which might preclude investigator participation if there is a monetary conflict of interest. For example, the American Association of Academic Medical Centers (AAMC) has suggested that a "rebuttable presumption" governs academics' conflicts of interest (AAMC 2003). Under this presumption, investigators who have a significant conflict of interest are presumed *not to* be allowed to do clinical studies until the presumption has been reviewed and reversed by an institutional conflicts of interest committee.

In the AAMC recommendations on conflicts of interest, we see the divergence of drug versus device regulation in stark relief. First is the question of execution. In a drug study, one does not need the expert biochemist to do a drug trial to see if a compound that worked in the lab will work in vivo. It is quite a different story with DBS when a team of investigators needs to apply selection criteria, choose suitable subjects, *and* then operatively insert and then program a device. That is a set of skills that remain at the forefront of investigational work and cannot effectively be delegated to others before proof of principle has been established (Fins and Schachter 2001; Fins 2008).

Second is the issue of discovery. In most drug studies the mechanism and the biology of the agent have been worked out before a trial has begun. It is quite the opposite in the context of DBS where the trial itself will reveal information which can be pivotal for both current and future studies. To regulate such trials as if they were solely exercises in optimizing therapeutics fails to contextualize them against the great possibilities of discovery of biological mechanism is to misconstrue their importance and potential (Fins 2012). Thomas Insel, director of the National Institute of Mental Health, said it best when he stressed the importance of understanding underlying mechanisms of brain-based disease and pathology and supplementing theoretical information with biological data. He presciently noted that "Despite the importance of wiring diagrams, [they] like the original genome maps, are necessary but not sufficient for understanding how the brain works" (Insel 2011). Insel's focus on basic biological mechanisms and their importance to further discovery cannot be stressed enough, but regrettably the current regulatory structure under which DBS studies are being conducted does not facilitate such discovery. More often than not, the regulatory schema is

a pathway to market and a roadblock to discovery. This shortcoming of regulation – and industry's response – can be best seen if we turn to the recent controversy over the FDA's approval of DBS in the obsessive-compulsive disorder under the *aegis* of the Humanitarian Device Exemption (HDE) (Fins et al. 2011c).

HDE for DBS in OCD: Regulatory Challenges

In February 2009 the FDA "approved" the use of DBS for the "treatment" of the obsessive-compulsive disorder (OCD) under the rubric of a Humanitarian Device Exemption (HDE) (FDA 2009, 2010a). The HDE approach was in lieu of the more conventional IDE (Investigational Device Exemption) route to market in which both safety and efficacy as demonstrated via a properly powered clinical trial must be demonstrated (Peña et al. 2007). HDEs operate via a lower standard of evidence requiring evidence of safety and do not require a clinical trial's evidence of efficacy for approval. This lower standard is meant to facilitate approval of devices that are intended for "orphan" markets where the disease, as verified by the FDA Office of Orphan Products Development, verifies that the condition annually affects 4,000 or fewer patients. If the disease meets these criteria, the device manufacturer can apply for the HDE from the Center for Devices and Radiological Health at the FDA (FDA 2010a, b). And if the approval is granted, FDA approval fees are waived.

Previously, along with colleagues from the Deep Brain Stimulation in Psychiatry Working Group of the *Europäische Akademie* and the *Universitätsklinikum Bonn*, I made the argument in *Health Affairs* that the FDA's approval of DBS for OCD was inappropriate (Fins et al. 2011a, c). Beyond contesting epidemiological data that the target population was less than 4,000 patients per year (Fins et al. 2011c; Kessler et al. 2005; Pallanti and Quercioli 2006; Hollander 1997), we asserted that the use of DBS in OCD was not amenable to the less stringent HDE pathway because the use of the stimulator was not merely an analogous use of a previously approved stimulator, which it was, but rather a novel application for the device with insertion into a new site, in a distinct disease condition. As such, we argued that it was a new application that required the use of the more stringent IDE application (Fins et al. 2011c).

And we made this argument for two reasons which illustrate my central thesis that approval of devices, as if they were pharmaceutical agents, will create logical misconstruals and that devices require their own, not as yet, fully articulated ethical and regulatory framework. First, in the FDA approval, there was a near exclusive focus on the device and not the context of its use. Second was the failure to appreciate that the manner of the device's approval had implications for therapy and research.

Let us start with the premise that because the device had been approved in one set of circumstances, such as its prior use in Parkinson's disease (CMS 2003), it should also be approved in the context of OCD. After all, the stimulator and its mechanics and electronics are known to be safe, so why not assert that the device was safe, much like one might assert that the pharmacology of a drug developed for

one purpose could be used off-label for another. We speculated that FDA's "familiarity with the device might have distracted them from seeing the novelty of the brain target and clinical indication as the key reason why an IDE was indicated as the preferred pathway to market" (Fins et al. 2011b).

That leap of logic from the world of drug development, pharmacology, and drug metabolism just does not work in the context of DBS where the functioning of the stimulator is but a part of the equation. Here, both safety and efficacy hinge on the placement of the electrode as a well as stimulation parameters which can themselves have variable impact based on settings and how those settings evoke biological effects in *a particular locale in the brain in a particular disease state*.

Approval by analogy to the use of the same device in a different context also fails to take account of what might be learned if the placement is done for a different reason, for a different disease. Beyond the therapeutic benefits which might accrue, basic knowledge might emerge from aggregating data from such implantations. Unfortunately, this is impossible when the placement is done under an HDE because one cannot generalize data from individualized cases implanted off protocol (Schlaepfer and Fins 2010). This is a lost opportunity for discovery and speaks to the FDA's regulation of DBS in OCD as if it were the regulation of a vetted therapy and not the investigational work that it remains (Fins et al. 2011c).

This confusion is revealed in FDA regulations concerning the HDE which requires that an HDE approval be reviewed by an Institutional Review Board (IRB), bodies charged with the oversight of research ethics within an institution in the United States. Yet paradoxically, even though IRB approval was required, FDA does not require that *research consent* be obtained. From their point of view, FDA views the insertion of a device under the aegis of an HDE as the practice of medicine (Hurley 2011), although it cannot explain why the practice of medicine should, in any fashion, be under the jurisdiction of an IRB (FDA 2006; Fins et al. 2011b).

FDA's HDE approval of Medtronic's "Reclaim" device for DBS in OCD (Medtronic 2009) was readily marketed by the company as a therapy even though no efficacy data had gone into that approval (Fins et al. 2011c). In advertising materials promulgated by Medtronic, the company "...announced today its official entrance into psychiatric therapies with approval from the U.S. Food and Drug Administration (FDA) for a humanitarian device exemption (HDE) for its Reclaim™ Deep Brain Stimulation (DBS) Therapy for chronic, severe obsessive-compulsive disorder (OCD).... Medtronic plans to make Reclaim DBS Therapy for OCD available in the United States by mid-2009" (Medtronic 2009; Dooren 2009). This was ethically problematic because it implies that a device is therapeutic prior to subjecting it to trials of efficacy, thus fostering the potential for a therapeutic misconception on the part of those who might receive the intervention (Lidz et al. 2004).

Economics, Market Scope, and a Proposed Remedy

Ultimately, the greatest consequence of regulating devices as if they were drugs is that the expectations of drug regulation are placed upon devices when the resources

required to achieve that degree of regulation are not in place. Given the aforementioned size of pharmaceutical houses as compared to device makers and the relative size of their respective markets, the expectation that a full-fledged IDE – the device analog of an IND (investigational new drug exemption) (FDA 2011) – can be mounted for novel DBS applications may be misplaced. According to my own experience with a DBS trial in the minimally conscious state (Fins and Schiff 2010; Schiff et al. 2007), and sources from within the neuropsychiatric DBS-investigative community, resources for an IDE trial are tenuous and very often nonexistent. This leads investigators to attempt to do their research in suboptimal ways, such as resorting to the HDE option.

The pressures of developing a highly capitalized device for a small market are captured in the practices of the Medtronic Corporation as noted by Daniel Bernard. In a review of how the company utilizes the humanitarian device exemption process to build upon its previously approved devices in order to develop – or in many cases rebrand – old tools for new purposes, Bernard (2009) reports how predicate devices approved for use in Parkinson's disease become the platform for additional devices, utilizing a "horizontal model" of product development. Bernard (2009) notes that "...an undisclosed source at Medtronic claimed that it would have been too costly and time-consuming to obtain a premarket approval for their Humanitarian Device Exemption-approved Activa Deep Brain Stimulation devices.... Medtronic saved 3 years, recovered $10 million from the initial research and development costs, and established a good relationship with many physicians within the field."

If DBS is both a clinical tool and a tool of discovery, it will be in a no-man's-land between commercial and governmental funding. It is quite possible that it will be abandoned by the market, which will see it as investigational and think it the responsibility of funders like the NIH. The NIH might look at its commercial success in the context of Parkinson's disease and view it as applied, not basic, science. In either scenario, the work will suffer.

So how might the regulatory system be reformed within budgetary constraints? In addition to reform of the Bayh-Dole Act and changes to the timing of IP exchange to broaden the scientific commons suggested elsewhere, (Fins 2010), another suggestion previously outlined would be a new regulatory pathway intermediate between the IDE and HDE, namely the *mini-IDE* (Fins et al. 2012). The mini-IDE would be able to address both safety *and* efficacy utilizing limited resources. In lieu of ill-fated attempts to use the HDE as a way to do (bad) science or enter the marketplace hamstrung without resources to enable work via the IDE pathway, the mini-IDE would seek to bridge the gap between bad science and good (or better) policy.

The mini-IDE would start with a predicate approval of a prior device, much as is currently used when an HDE application is being vetted (or like the 510(k) approval process used in lower risk Class II devices (Committee Public Health 2011)). This approval based on safety data would be necessary but not sufficient and would therefore result in *contingent* licensure. This would need to be maintained over time meeting progressive standards approaching the more rigorous IDE. To move from contingent licensure to full approval, the device and its target would be subject to participation in a standardized hypothesis-driven trial as in the IDE process.

Mini-IDE data would however be generated differently than that obtained from the generic IDE process. Instead of collecting all the data in a single, time-limited, and capital-intensive trial, data for the mini-IDE study would be disseminated over time and space. The data would be generated over time under the rubric of contingent licensure and could take years to accrue to meet a statistical standard of efficacy. And once that standard was reached, the mini-IDE trial would graduate, as it were, and have the data necessary to be viewed as an IDE, allowing for approval under that more rigorous pathway.

The study would also be geographically dispersed and performed by different investigators, one at a time, much like current practice with the HDE. The key difference would be that to be eligible for the required contingent, mini-IDE licensure, investigators would have to agree to abide by a single centralized protocol that would allow for the pooling of results, outcome data, and adverse events. Data and a data-safety monitoring board would be centralized – we previously proposed that this might be a role that the Foundation for the NIH (FNIH) might play (Fins et al. 2012) and that a study registry be created to allow for data mining and the development of properly powered scientific analyses of safety and efficacy (Schlaepfer and Fins 2010; Synofzik et al. 2011).

The advantage of the mini-IDE schema is that it would allow hypothesis-driven work to proceed with limited resources and allow for the collection of aggregate data and the pooling of adverse event information which might have material value when viewed outside of the context of individual patient "studies." Pooled analysis would make both implicit toxicities more evident and potentially suggest novel therapies based on unanticipated "side effects."

Fiscally, the dispersion of the trial in time and space amortizes the cost. It would allow work to proceed without all the upfront capital expenditures that are necessary for an IDE. It would also maintain access to the benefits of studies like the HDE while providing the added benefit that study participation would result in generalizable data, something that would be impossible under the HDE approval process. And as importantly, the contingent licensure reminds the patient and the investigator that the device is still not a fully vetted therapy and thus helps to mitigate the challenge of the therapeutic misconception.

DBS as Personalized Medicine

Earlier, it was asserted that the placement of a deep brain stimulator was more akin to personalized medicine than conventional drug development. Instead of drugs produced for the masses, the insertion of an electrode into a particular brain in a highly personal matter must take account one's anatomy, physiology, and in the case of more experimental applications, as in traumatic brain injury, the lesions of an injured brain. For example, in Parkinson's disease, the target is chosen very individually. Sometimes it is the subthalamic nucleus (STN), the globus pallidus interna (GPi), and sometimes (in subjects with freezing of gait) the STN plus the substantia nigra pars reticulata (SNr).

In this way, this emerging therapeutic is much like personalized medicine where a drug is tailored, indeed developed to the illness of a *particular* patient. It is highly particularized and personal, targeted, and precision therapy designed to correct, override, or protect the individual against a deranged DNA sequence somewhere in the patient's genome. The ethical and regulatory challenges posed by this approach to therapy will not be fully articulated until we have more experience with personalized medicine (Soden et al. 2012). Nonetheless, some commentators have begun to point to the tension raised by drugs developed for individuals and not for the masses (Maglo 2012), a line of argument which echoes the dynamic between the regulation of drugs for the many and devices for the few.

That the same problem is now being confronted within the sphere of "conventional" versus "personalized" medicine suggests that the discordance between the FDA's regulatory approach to drugs versus devices is reemerging in a different and perhaps more powerful guise. And given the therapeutic power of this approach, regulatory issues will need to be identified and addressed.

As drugs become more personalized in design and scope, they will begin to resemble the nature of interventions like deep brain stimulation which have been, as noted, poorly regulated through the explantation of methods better geared to the oversight of mass-produced drugs than tailored therapies. Like devices, approaches to personalized medicine will be costly, personalized, and often idiosyncratic providing a therapy to some and knowledge of biological mechanism to others. Others might benefit from the new knowledge of a biochemical pathway elucidated through personalized medicine much as patients with brain injury are benefiting from the identification of the *mesocircuit* as a substrate of consciousness (Schiff and Posner 2007; Schiff 2010), a discovery made possible – in part – by work utilizing deep brain stimulation in the minimally conscious state (Fins 2012).

This generalization of knowledge from N = 1 cases shows that the personalized approach may be either therapeutic or probative. It also illustrates that a new method of oversight must come to the fore that allows discovery to proceed even as therapeutics emerge. As we saw earlier, the Humanitarian Device Exemption did not and could not achieve these dual goals. Novel approaches, such as the suggested mini-IDE, might prove an interim remedy to the challenges encountered in regulating and – as importantly – sustaining discovery through DBS.

Conclusion

The advent of molecular personalized medicine, to complement practices which have long existed in deep brain stimulation, will help motivate regulators to formulate new approaches to oversee the development of treatments for the few which might benefit the many. Now that DBS has an analog in personalized medicine, the problem of proportionate regulation becomes a shared one. It is an

urgent challenge to which all must attend, if progress is to be sustained in all realms of personalized medicine.

Acknowledgement/Disclosures Dr. Fins acknowledges the collegiality of the working group, "Deep Brain Stimulation in Psychiatry. Guidance for Responsible Research and Application," of the *Europäische Akademie. Universitätsklinikum Bonn*. Bonn, Germany. He also gratefully notes the support of the Buster Foundation, the Jerold B. Katz Foundation, as well as a Clinical and Translational Science Center (UL1)-Cooperative Agreement (CTSC) 1UL1 RR024996 to Weill Cornell Medical College and its Ethics Core. He also thanks the editors of this volume for their invitation to collaborate and their helpful critiques.

References

AAMC Task Force on Financial Conflicts of Interest in Clinical Research. (2003). Protecting subjects, preserving trust, promoting progress I: Policy and guidelines for the oversight of individual financial interests in human subjects research. *Academic Medicine, 78*(2), 225–236.

Abelson, R. (2007, October 27). Medtronic, again questioned over payments to doctors, Is subject of Senator's inquiry. *The New York Times*. Retrieved from http://www.nytimes.com/2007/09/27/business/27letter.html?scp=25&sq=&st=nyt

Armstrong, D. (2008, September 25). Lawsuit says Medtronic gave doctors array of perks. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB122230535985873827.html

Bayh-Dole Act (P.L. 96-517, Patent and Trademark Act Amendments of 1980). 37 C.F.R. 401 and 35 U.S.C. 200-212.

Bernad, D. M. (2009). Humanitarian use device and humanitarian device exemption regulatory programs: Pros and cons. *Expert Review of Medical Devices, 6*(2), 137–145.

Bronstein, J. M., Tagliati, M., Alterman, R. L., Lozano, A. M., Volkmann, J., Stefani, A., Horak, F. B., Okun, M. S., Foote, K. D., Krack, P., Pahwa, R., Henderson, J. M., Hariz, M. I., Bakay, R. A., Rezai, A., Marks, W. J., Jr., Moro, E., Vitek, J. L., Weaver, F. M., Gross, R. E., & Delong, M. R. (2011). Deep brain stimulation for Parkinson disease: An expert consensus and review of key issues. *Archives of Neurology, 68*(2), 165. doi:10.1001/archneurol.2010.260.

Carpenter, D. (2010). *Reputation and power: Organizational image and pharmaceutical regulation at the FDA*. Princeton: Princeton University Press.

Centers for Medicare and Medicaid Services and Department for Health and Human Services. (2003, February 14). *Deep brain stimulation for essential tremor and Parkinson's disease*. Change Request 2553. Retrieved from http://www.cms.gov/transmittals/downloads/AB03023.pdf. Accessed 10 Dec 2010.

Committee on the Public Health Effectiveness of the FDA 510(k) Clearance Process; Institute of Medicine. (2011). *Medical devices and the public's health: The FDA 510(k) clearance process at 35 years*. Washington, DC: The National Academies Press.

Dooren, J. C. (2009, February 19). FDA approves Medtronic brain device to treat severe cases of OCD. *Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB123507654820526043.html

Fins, J. J. (1995). The hospital as ecosystem. *Ecosystem Health, 1*(4), 255–259.

Fins, J. J. (2004). Deep brain stimulation. In S. G. Post (Ed.), *Encyclopedia of bioethics* (3rd ed., Vol. 2, pp. 629–634). New York: Macmillan.

Fins, J. J. (2007). Disclose and justify: Intellectual property, conflicts of interest, and neurosurgery. *Congress Quarterly (The Official Newsmagazine of the Congress of Neurological Surgeons), 8*(3), 34–36.

Fins, J. J. (2008). Surgical innovation and ethical dilemmas: Precautions & proximity. *Cleveland Clinic Journal of Medicine, 75*(Suppl 6), S7–S12.
Fins, J. J. (2010). Deep brain stimulation, free markets and the scientific commons: Is it time to revisit the Bayh-Dole Act of 1980? *Neuromodulation, 13*, 153–159.
Fins, J. J. (2012). Deep brain stimulation as a probative biology: Scientific inquiry & the mosaic device. *American Journal of Bioethics-Neuro Science, 3*(1), 4–8.
Fins, J. J., & Schachter, M. (2001). Investigators, industry and the heuristic device. Ethics, patent law and clinical innovation. *Accountability in Research, 8*(3), 219–233.
Fins, J. J., & Schiff, N. D. (2010). Conflicts of interest in deep brain stimulation research and the ethics of transparency. *The Journal of Clinical Ethics, 21*(2), 125–132.
Fins, J. J., Rezai, A. R., & Greenberg, B. D. (2006). Psychosurgery: Avoiding an ethical redux while advancing a therapeutic future. *Neurosurgery, 59*(4), 713–716.
Fins, J. J., Mayberg, H., & Schlaepfer, T. E. (2011a). Humanitarian device exemptions: The authors' reply. *Health Affairs (Millwood), 30*(6), 1213.
Fins, J. J., Mayberg, H., & Schlaepfer, T. E. (2011b). FDA exemptions: The authors' reply. *Health Affairs (Millwood), 30*(6), 1212.
Fins, J. J., Mayberg, H. S., Nuttin, B., Kubu, C. S., Galert, T., Strum, V., Stoppenbrink, K., Merkel, R., & Schlaepfer, T. (2011c). Neuropsychiatric deep brain stimulation research and the misuse of the humanitarian device exemption. *Health Affairs, 30*(2), 302–311. doi:10.1377/hlthaff.2010.0157.
Fins, J. J., Schlaepfer, T. E., Nuttin, B., Kubu, C. S., Galert, T., Sturm, V., Merkel, R., & Mayberg, H. S. (2011d). Ethical guidance for the management of conflicts of interest for researchers, engineers and clinicians engaged in the development of therapeutic deep brain stimulation. *Journal of Neural Engineering, 8*(3), 033001. doi:10.1088/1741-2560/8/3/03301.
Fins, J. J., Dorfman, G. S., & Pancrazio, J. J. (2012). Challenges to deep brain stimulation: A pragmatic response to ethical, fiscal and regulatory concerns. Proceedings from "deep brain stimulation." 91st annual conference of the association for research in nervous and mental disease. *Annals of the New York Academy of Sciences, 1265*, 80–90. doi: 10.1111/j.1749-6632.2012.06598.x.
Food and Drug Administration. (2006). *Information sheet guidance for IRBs, clinical investigators, and sponsors: Frequently asked questions about medical devices*. Resource document. Retrieved from http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM127067.pdf. Accessed 10 Dec 2010.
Food and Drug Administration. (2009). *Approval order H05003. Letter to Patrick L. Johnson, Medtronic Neuromodulation from Donna-Bea Tillman, Ph.D, M.P.A., Director, Office of Device Evaluation, Center for Devices and Radiologic Health, FDA*. Resource document. Retrieved from http://wwwaccessdatafdagov/cdrh_docs/pdf5/H050003apdf
Food and Drug Administration. (2010a). *Guidance for HDE holders, Institutional Review Boards (IRBs), clinical investigators, and FDA staff—Humanitarian Device Exemption (HDE) regulation: questions and answers*. Resource document. Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm110194.htm. Accessed 10 Dec 2010.
Food and Drug Administration. (2010b). *Office of orphan product development*. Website. Retrieved from http://www.fda.gov/AboutFDA/CentersOffices/OC/OfficeofScienceandHealthCoordination/OfficeofOrphanProductDevelopment/default.htm. Accessed 10 Dec 2010.
Food and Drug Administration. (2011). *Investigational new drug (IND) application*. Application. Retrieved from http://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/investigationalnewdrugindapplication/default.htm. Accessed 28 Nov 2012.
Hollander, E. (1997). Obsessive-compulsive disorder: The hidden epidemic. *The Journal of Clinical Psychiatry, 58*(Suppl 12), 3–6.
Hurley, D. (2011). Should the FDA rescind the humanitarian exemption for DBS? *Neurology Today, 11*(5), 10.

Insel, T. (2011). *Neuroscience advances showcased in Washington*. National Institute of Mental Health Directors Blog. Retrieved from http://www.nimh.nih.gov/about/director/index.shtml

Johnson & Johnson. (2011). *Stock information*. http://www.investor.jnj.com/stock-information.cfm. Accessed 22 Nov 2012.

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 617–627.

Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Zumsteg, D., Wennberg, R., Keren, R., Wherrett, J., Naglie, G., Hamani, C., Smith, G. S., & Lozano, A. M. (2010). A phase I trial of deep brain stimulation of memory circuits in Alzheimer's disease. *Annals of Neurology, 68*, 521–534.

Lidz, C. W., Appelbaum, P. S., Grisso, T., & Renaud, M. (2004). Therapeutic misconception and the appreciation of risks in clinical trials. *Social Science & Medicine, 58*(9), 1689–1697.

Maglo, K. N. (2012). Group-based and personalized care in an age of genomic and evidence-based medicine: A reappraisal. *Perspectives in Biology and Medicine, 55*(1), 137–154.

Medtronic. (2009, February 19). *Medtronic receives FDA HDE approval to commercialize the first deep brain stimulation (DBS) therapy for a psychiatric indication in the United States*. News Release. Retrieved from http://wwwp.medtronic.com/Newsroom/NewsReleaseDetails.do?itemId=1235065362795&lang=en_US. Accessed 13 Dec 2010.

Pallanti, S., & Quercioli, L. (2006). Treatment-refractory obsessive-compulsive disorder: Methodological issues, operational definitions and therapeutic lines. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 30*(3), 400–412.

Peña, C., Bowsher, K., Costello, A., De Luca, R., Doll, S., Li, K., et al. (2007). An overview of FDA medical device regulation as it relates to deep brain stimulation devices. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 15*(3), 421–424.

Pressman, J. D. (1998). *Last resort: Psychosurgery and the limits of medicine*. New York: Cambridge University Press.

Reuters. *Sources for financial data and market capitalization*. http://www.reuters.com. Accessed 22 Nov 2012.

Schiff, N. D. (2010). Recovery of consciousness after brain injury: A mesocircuit hypothesis. *Trends in Neuroscience, 33*(1), 1–9.

Schiff, N. D., & Posner, J. B. (2007). Another "awakenings". *Annals of Neurology, 62*(1), 5–7.

Schiff, N. D., Giacino, J. T., Kalmar, K., Victor, J. D., Baker, K., Gerber, M., Fritz, B., Eisenberg, B., O'Connor, J., Kobylarz, E. J., Farris, S., Machado, A., McCagg, C., Plum, F., Fins, J. J., & Rezai, A. R. (2007). Behavioral improvements with thalamic stimulation after severe traumatic brain injury. *Nature, 448*(7153), 600–603.

Schlaepfer, T. E., & Fins, J. J. (2010). Deep brain stimulation and the neuroethics of responsible publishing: When one is not enough. *Journal of the American Medical Association, 303*(8), 775–776.

Schuepbach, W. M. M., et al. (2013). Neurostimulation for Parkinson's disease with early motor complications. *The New England Journal of Medicine, 368*, 610–622.

Soden, S. E., Farrow, E. G., Saunders, C. J., & Lantos, J. D. (2012). Genomic medicine: Evolving science, evolving ethics. *Personalized Medicine, 9*(5), 523–5238.

Synofzik, M., & Schlaepfer, T. E. (2011). Electrodes in the brain—ethical criteria for research with deep brain stimulation for neuropsychiatric disorders. *Brain Stimulation, 4*, 7–16.

Synofzik, M., Fins, J. J., & Schlaepfer, T. E. (2011). A neuromodulation experience registry for deep brain stimulation studies in psychiatric research: A rationale and recommendations for implementation. *Brain Stimulation, 5*(4), 653–655.

Valldeoriola, F., Morsi, O., Tolosa, E., Rumià, J., Martí, M. J., & Martínez-Martín, P. (2007). Prospective comparative study on cost-effectiveness of subthalamic stimulation and best medical treatment in advanced Parkinson's disease. *Movement Disorders, 22*(15), 2183–2191. doi:10.1111/j.1525-1403.2009.00238.x.

Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations

38

Matthis Synofzik

Contents

Abstract

The applications of deep brain stimulation (DBS) are rapidly increasing and now include a large variety of neurological and psychiatric diseases, such as addiction, Alzheimer's disease, anorexia nervosa, and rare movement disorders. High-quality data about benefit and harm of DBS in these disorders remain rare as many findings rely on small cohorts or single cases, variable methodologies, and differing outcome measures. Collectively, these problems make the field prone to bias and selective reporting, evoke ethical concerns regarding possibly premature expansions to new conditions without appropriate justification and research, and indicate the possibility that media, the public, and institutional review boards might be easily misguided by some reports. Thus, these problems are primarily not of scientific or methodological nature, but of *ethical* nature. Here, three approaches are suggested on how these problems might be reduced: by an optimization of trial designs, the implementation of

M. Synofzik
Department of Neurology, Centre for Neurology and Hertie-Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany
e-mail: matthis.synofzik@uni-tuebingen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_32,

standards of reporting, and the creation of a DBS study register which includes in particular single-case studies or case series. Future work has to work out these proposals in more detail and study its effectiveness when implemented in practice.

Introduction

Randomized clinical trials (RCTs) and clinical experience have demonstrated that deep brain stimulation (DBS) is a powerful tool in the treatment of movement disorders like Parkinson's Disease (PD), Essential Tremor (ET), or dystonia. In the last years, these observations were complemented by promising results from small cohort studies of DBS in treatment-refractory patients with neuropsychiatric diseases like obsessive-compulsive disorder (OCD), major depression (MD), or Tourette's syndrome. Based on these experiences, the applications of DBS are now rapidly spreading all over the world to novel applications in patients with other neuropsychiatric diseases or rare movement disorders. There are first preliminary results reporting an effectiveness of DBS in case series of patients with anorexia nervosa (Lipsman et al. 2013), obesity, violent behavior (Franzini et al. 2005), minimal conscious state (MCS) (Schiff et al. 2007), alcohol addiction (Muller et al. 2013), Alzheimer's Disease (Laxton et al. 2010), or rare movement disorders (e.g., chorea acanthocytosis (Li et al. 2012), ataxia with abetalipoproteinemia (Mammis et al. 2012), or neurodegeneration with brain iron accumulation (Castelnau et al. 2005)) (for an overview, see (Hariz et al. 2013)). This rapid spread evokes strong ethical concerns regarding the possibility of premature expansion to new conditions without appropriate justification and research. Moreover, it might be seen as in indirect indicator of a widespread acceptance and usage of neuropsychiatric and movement disorder DBS before adequate safety and efficacy data are obtained. This possibility is worrisome in particular as DBS is not only rapidly moving toward novel indications, but also from larger specialized academic hospitals to smaller centers and even private practice (Bell et al. 2009).

In the early phase of its scientific investigation and clinical application, each novel DBS application is associated with the following problems:

– Limited evidence about the effectiveness of DBS (as RCTs will be lacking in the early stages of investigation)
– Small subject numbers, including only a single or a few patients
– Use of different outcome measures and variable methods of assessment within and between academic centers
– Reports with largely varying degrees of methodological and scientific quality

These problems are certainly not specific to DBS research, but apply to research in many other novel biotechnological domains as well (Clausen 2009). Thus, many of the ethical approaches already developed in the field of research ethics might be applicable to DBS research as well (for a general overview, see ▶ Chap. 70, "Human Brain Research and Ethics"), yet they need a DBS-specific articulation and specification (Clausen 2010). This is urgently needed in the field of DBS

since – despite the aforementioned problems – there is a tendency to rapidly accept apparently positive results from novel DBS applications and extend its usage. This can currently be observed with respect to DBS in some rare movement disorders or neuropsychiatric diseases (e.g., substance addiction) (Synofzik and Schlaepfer 2011). Importantly, the aforementioned problems are not only of *scientific* or *methodological*, but also of *ethical* nature.

– These problems a priorily constrain the possibility to validly assess the chances of benefit and risks of harm of a novel DBS application. This, however, is inevitably needed to determine the ethical criteria of beneficence and maleficence, which should be part of each ethical evaluation of DBS (Synofzik et al. 2012; Synofzik and Schlaepfer 2008, 2010).
– The excessive reliance on single-patient or small case series reports makes these nascent DBS domain highly vulnerable for selective reporting (Schlaepfer and Fins 2010; Synofzik et al. 2012). It opens up the possibility that only positive results (with often insufficiently controlled study designs) will be published, while negative data that might also have important implications will not be brought to the public (Schlaepfer and Fins 2010). Due to the inclusion of only one or few subjects, such small negative studies can easily be performed and analyzed without further attention by other researchers within and outside the respective centers. This well-known problem of selective reporting is already indirectly evidenced in the young field of neuropsychiatric DBS: several single-case studies have been published only because of interesting secondary effects, whereas the primary outcome effects were not achieved (Schlaepfer and Fins 2010). This indirectly indicates that there might be many other unpublished single-case DBS interventions for which the primary outcomes were not achieved and no interesting secondary effects were observed. Such a selective reporting will lead to a severe distortion of available evidence in this field which might harm future patients with neuropsychiatric disease or rare movement disorders (Dickersin and Rennie 2003).
– The above mentioned problems increase the likelihood that the reported scientific results are false positive or false negative with respect to the real outcomes. This will misguide other researchers, the media, and the public.
– They constrain the decision-making process of institutional review boards (IRBs) on the legitimacy of novel DBS studies and applications. Here, IRBs do not have a sufficient evidence basis to judge about the potential benefits and harms of a planned study. Moreover, IRBs might have to rely on potentially misguiding single-case reports and false positive or false negative small cohort studies for their decisions.

Thus, it is primarily not (only) a scientific, but an *ethical* necessity to call for a more rigorous methodological quality of DBS studies, comprehensive standards of DBS outcome reporting, prevention of selective reporting, and safeguards constraining a premature spread of DBS to novel applications and less experienced centers. This is of particular importance in the rapidly emerging field of DBS where many novel applications include children and/or complex neurodegenerative and neuropsychiatric diseases. Despite these obvious ethical needs, investigations

primarily focusing on the ethics of DBS research and reporting are still rare. Here suggestions for optimal trial designs, standardized reporting, and study registrations are outlined which might reduce the ethical problems outlined above.

Optimal Trial Designs

What level of evidence should be optimally used to establish safety and efficacy before a novel DBS application can move forward to clinical application? This complex question cannot be addressed here in full detail, but from an ethical perspective, it seems obvious (and almost trivial) to ask for randomized controlled trials (RCTs) as only they can establish true efficacy and rule out possible confounding factors. Depending on the expected effect size, RCTs are possible even for cohorts of <20 subjects. Even if a standard RCT would not be possible due to inadequate power, for example, in rare movement disorders, there is still a large variety of different trial designs that can be used to yield a randomized, comparative clinical trial (Cornu et al. 2013).

It is more difficult to determine the ethically most appropriate control condition. An active-controlled study design resorting only to best medical treatment (i.e., pharmacotherapy and/or behavioral therapy) does not completely suffice, as it could not rule out placebo responses to DBS insertion and/or DBS programming and would thus not allow making a final decision about the quality of DBS. An active placebo-controlled design, resorting to sham-stimulation (e.g., 0.0 V or subthreshold stimulation) and comprising, interalia, of surgery, electrode insertion, and parameter setting, seems to be the only way to control for such placebo effects (for an overview on sham surgery, see ▶ Chap. 72, "Ethics of Sham Surgery in Clinical Trials for Neurologic Disease"). Requesting such an optimally controlled study design is not just a scientific, but also an ethical demand: if DBS trials would refrain from using active placebo, DBS treatments might be approved, which appear more effective than standard treatments, but are, in fact, no more effective than placebo (in this case, sham-stimulation). This would lead to unjustified risk of harm for patients treated with DBS and to high, unnecessary intervention costs for the health care system. The feasibility of DBS sham-stimulation has already been demonstrated in a large-scale study in dystonia patients (Kupsch et al. 2006), in a smaller study in OCD patients (Mallet et al. 2008) and was also exemplarily investigated in patients with depression using off-on-off-on trials of varying short periods as well as a blinded discontinuation phase of 4 weeks (Mayberg et al. 2005; Schlaepfer et al. 2008). After a limited, well-defined study period, patients from the sham-stimulation arm should receive active stimulation.

This trial design is almost ethically innocuous, because patients are not asked to forego, but only to delay treatment of presumed benefit. Thus, the chance of presumed benefit might equiponderate the nontrivial risk of harm of DBS surgery (Synofzik and Schlaepfer 2011), which these patients have gone through. This design also prevents exploitation of the placebo-study arm for research purposes, because patients' participation confers presumed benefits not only to future patients

and society from generating biomedical knowledge (beneficence in clinical research) but also to the individual patients themselves (beneficence in clinical medicine). Importantly, the shift between sham-stimulation and active stimulation does not necessarily need to be performed within a double crossover design (AB/BA design) as this would mean that patients might experience major exacerbation of the underlying disease if effective active stimulation was substituted by ineffective sham-stimulation. In fact, a study of DBS in major depression has demonstrated that an AB/BA design might lead to such ethically highly problematic consequences and, subsequently, its cessation (Bewernick et al. 2010). Thus, it might suffice if only one study arm crosses from sham-stimulation to active stimulation (AA/BA design), i.e., the trial would follow a randomized controlled study design with *delayed start* of the control group.

Standardized Reporting

As already proposed earlier (Synofzik and Schlaepfer 2011), guidelines are needed not only for optimal *application* of DBS (including presurgical assessment, surgery, and long-term monitoring of DBS), they are also needed for ethically sound *publishing* in the field of DBS. Standards and guidelines of scientific reporting will lead to an improved quality of reporting (Coultas 2007). This, in turn, will lead to a better communication of the results to the public and the media as well as to a better evidence basis from which beneficence and maleficence can be evaluated within ethical analyses by IRBs, researchers, and clinicians.

In addition to the general standards of scientific reporting (CONSORT Standards [Consolidated Standards of Reporting Trials] (Moher et al. 2001)), the standards for DBS reporting should include several additional specific aspects.

1. The reported outcome parameters should include a detailed description also of the (long-term) side effects. It is now known from studies of long-term effects of DBS in both Parkinson's Disease (see ▶ Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects") and Essential Tremor that DBS might be effective in treating some motor features, yet that this "comes with a price" (Warnke 2013): even in experienced hands, it comes with a range of side effects like worsening of verbal fluency and development of gait ataxia and dysarthria, respectively (Warnke 2013). These side effects will only come to the fore if they are searched for in detail and if long-term effects are responsibly reported in the public.

   Moreover, the reported outcome parameters should include not only symptom-specific or functional outcomes (e.g., changes in a disease-specific symptom score), but also outcomes that are meaningful in the individual daily life of the patients. This call for more comprehensive outcome reporting applies not only to DBS study designs per se, but also to follow-up assessments. They should integrate ***broader biopsychosocial outcome measures*** like reintegration into family life, social and work environments, and psychosocial and global quality of life. Moreover, to further improve assessment whether observed DBS

outcomes are truly beneficial or not, they should be complemented by ***qualitative measures***. This is necessary as changes in disease experiences, coping strategies, and psychosocial life situations are hard to capture by current (semi-)quantitative scales.

The need for such measures is illustrated by a study in PD patients (Agid et al. 2006; Schupbach et al. 2006): in spite of, or probably because of a clear improvement in various outcome variables after DBS implantation, many are not happier with their lives. They go through tormented periods in their marriages or fail to resume professional activities postoperatively. This experience of PD patients does not seem to be specific to that condition, but can be expected after rapid symptom modification in any chronic life-determining disease, psychiatric or somatic. For example, a rapid improvement of major depression, obsessive-compulsive disorder, or addiction will change previous personal activities of daily living and social relations in a more drastic and rapid way than can be compensated for by the individual himself/herself and by his personal surroundings. It was reported for a patient with PD and for a patient with MD that after "highly successful" DBS treatments, comorbid personality disorders came to the fore (Schlaepfer and Fins 2010). This effect seriously complicated the following treatment and reduced quality of life measures below the pre-DBS state (Schlaepfer and Fins 2010). These different experiences from different centers and different patient groups clearly indicate that merely reporting symptom-specific scores can be highly misleading with respect to the true net benefit for the patient. This can only be captured by more comprehensive outcome reporting.

The demand to report individually meaningful outcomes of daily living and quality of life beyond mere symptom-specific scores can already be applied to current studies. For example, in a trial of seemingly "successful" DBS for Minimal Conscious State (MCS), one might cautiously ask about the potential for incremental distress from increased self-awareness associated with cognitive improvement (Fins 2000). Even if the MCS patients would not suffer from increased mental distress, it remains questionable whether the improvements in vigilance are sufficient to lead to an improvement in quality of life.

The quality of life outcomes assessed within a study should not only be reported in the appendix, but in the main text of a report. A recent DBS case series of patients with anorexia nervosa (Lipsman et al. 2013) reported the quality of life outcomes only in the appendix. Interestingly, in this series, three out of six patients did not show improvements in quality of life (cf. de Zwaan and Schlaepfer 2013).

2. The scientific reports have to systematically report also ***all nonsurgical and nondrug therapies*** received by the study patients, both before and after DBS surgery. Any changes in these therapies might essentially contribute to the observed effects which might be prematurely related to DBS, in particular in neuropsychiatric diseases. For example, the additional care by a psychotherapist or its withdrawal, a change in the person of the treating psychotherapist, or a novel interdisciplinary counseling or novel expert therapy might lead to substantial effects that are not related to DBS, but, for example, to the

preoperative assessment in the course of DBS or simply to the change of the patient from an outside clinics and physician to a tertiary care center. If these preceding and subsequent therapies and the components of the therapies and the treatment team are not systematically captured and reported by the study investigators, the reported study effects are at risk to be (falsely) associated with the DBS insertion (in the intervention group) or non-DBS insertion (in the control group). This will lead to false positive or false negative results. For example, a phase 1 study on DBS in anorexia nervosa that was recently published in a high rank journal (Lipsman et al. 2013) did not report these therapies and the elements of the treatment team, let alone their continuation and discontinuation, respectively. This hampers, and potentially confounds, the interpretation of the overall results as well as single results. For example, it makes it hard to explain why subjects still gained more weight before surgery (2.4 body mass index [BMI] points) than they did afterward (0.5 BMI points), whether their postoperative weight loss (mean 2 BMI points) was really due to loss of enhanced in-patient care rather than an adverse effect of DBS surgery, and whether the eventual postoperative recovery from weight loss was really due to a DBS effect rather than increased intensive support after postsurgery weight loss (Hutton 2013).

Likewise, another case series of DBS in patients with anorexia nervosa (Wu et al. 2013) did not mention whether psychotherapy, which is considered the treatment of choice, was performed. This indirectly suggests that it was not applied at all. Thus, not reporting the nonsurgical therapies makes it difficult to decide whether the study did appropriately select and treat the subjects or not. If subjects did not receive the treatment of choice prior to DBS, they might have been undertreated prior to DBS. Moreover, if it is not reported whether they did or did not receive treatments according to the current guidelines (an information that was also omitted in the report by Wu and colleagues (Wu et al. 2013)), it cannot be decided if the subjects were adequately treated prior to DBS or not (de Zwaan and Schlaepfer 2013). It should also be reported whether or not the subjects received guideline-based treatments for their disease *after* the surgery (de Zwaan and Schlaepfer 2013). This fact – which was also not reported, for example, in the DBS study for anorexia nervosa by Lipsman et al. (2013) – might change the outcomes of the study. In addition, omission of a guideline-based treatment might also need particular ethical justification.

3. Scientific reports have to systematically include information about all the consequences of DBS in the psychosocial domain, in particular also in those ***patients where DBS was not effective***. This is important as, for example, in novel neuropsychiatric DBS applications like MD or OCD, a large share of patients does not respond to DBS – in several studies up to 50 %. These patients are often not of any further interest for the study investigators, whereas the patients where DBS was effective receive many further up assessments, counseling, and in-patient care within and outside of study protocols. Thus, it has to be reported whether and how ***follow-up care for DBS-negative patients*** was ensured, and which alternative care and treatment options (instead of DBS) were offered by the specialist center.

4. Scientific reports have to explicitly discuss whether the reported results are not only statistically significant, but indeed ***meaningful in terms of further treatment*** of the patients. Moreover, they should explicitly discuss the possibility whether they might lead to misinterpretation and possible misguidance of IRBs, media, and the public. Several recent uncontrolled observational DBS studies on single cases or small case series might have been correct in themselves; however, they remain vague and partly dubious about the question which consequences could by validly drawn for clinical practice. Thus, they opened the possibility to misguide patients, the public, and IRBs that have to judge on the legitimacy and adequacy of similar studies and of clinical application of DBS in this context. For example, a study on ten patients with DBS of the nucleus accumbens showed that three out of ten patients stopped smoking in the further course of the DBS treatment (Kuhn et al. 2009). A control group that would have received DBS at a different brain site, or surgery elsewhere at the body (e.g., in the abdomen), was missing. Thus, it remained unclear whether the (already relatively small) observed effect was a specific effect of DBS of the nucleus accumbens or might be expected also by DBS of other brain sites or even by any larger surgery with intensive presurgery and postsurgery care, and might thus be completely unspecific. The public, media, and IRBs might be prematurely inclined to accept the assumption that DBS of the nucleus accumbens would present a promising option in the treatment of tobacco addiction. Evidence for premature misinterpretations of study results by the media can be drawn, for example, from the report of DBS in anorexia nervosa (Lipsman et al. 2013). Even though the authors performed a careful interpretation of their study results, namely, primarily in terms of safety of the procedure, the general press interpreted the results very differently (de Zwaan and Schlaepfer 2013). Such a bias in media-reporting will be facilitated even more if researchers themselves do not discuss more explicitly whether or not their study results already imply clinical effectiveness, and if they do not explicitly discuss potential misinterpretations that might mislead the media and the public.

Study Registration

Even if DBS studies would follow an optimal trial design and the standards of outcome reporting, one problem remains: the current excessive reliance on single-patient or small case series in the field of neuropsychiatric DBS or of DBS for rare movement disorders. As outlined above, this trend makes the DBS domain highly vulnerable for selective reporting. It is of upmost importance of this nascent domain to know and share both positive and negative results of isolated studies and observations that would otherwise not be reported by the broader scientific community. Moreover, this trend leads to the fact that an important chance to share and aggregate data on efficacy and safety of DBS is still neglected.

To encounter these problems, a central registry for DBS was proposed (Synofzik et al. 2012). This registry is characterized by one important specific feature: in

contrast to a prototype registry like *Clinicaltrials.gov*, which does not include small feasibility and pilot studies of devices as well as single-case interventions, the eligibility of the DBS registry should not be limited only to rigorous clinical trials. Instead, in addition to established clinical trials, this registry should also include single-case DBS studies, clinical reports without control or comparison group, and isolated cases using human device exemptions (HDEs). Both interventional and observational studies would be accepted and trials should be registered irrespective of the anticipated or hypothesized outcome.

In addition to this prospective enrollment, retrospective clinical observations about DBS effects in neuropsychiatric patients can also be posted. This applies, for example, to relevant clinical or surgical anecdotes, important stimulation parameter findings (be they beneficial or detrimental) or interesting side effects (e.g., perioperative mood effects (Okun et al. 2007) or enhanced memory functioning (Hamani et al. 2008)).

But is such an effort really necessary? As described before (Synofzik et al. 2012), the key arguments supporting implementation of a specific DBS registry are mainly centered on patient safety and trial efficacy, yet they also serve many additional goals such as serving as a platform for data aggregation, regulatory oversight, and dissemination of innovative techniques (see Box 1).

For example, meta-analysis of pooled aggregate data from multiple small trials linked in a registry may demonstrate heretofore unrecognized efficacy in the fields of neuropsychiatric DBS or rare movement disorder DBS which are still marked by single-case reports and small cohort studies. Rather than a single case, it will be the body of evidence, aggregating many trials, that will have a persisting impact on medical practice (DeAngelis et al. 2004). Also information about adverse events can be aggregated and analyzed which might not be collected otherwise. This is of particular importance as it would not be possible for a patient or an IRB to assess the risks of participation in a DBS trial if an unknown proportion of adverse effects on the proposed intervention were not registered and not publicly available

Box1: Advantages of a DBS Registry in Neuropsychiatric DBS

- Accumulate beneficial effects
- Aggregate adverse events
- Collect long-term data
- Collect data on ineffective DBS/failed trials
- Give back the benefit of aggregated information to the patient
- Identify side effects that might serve as therapeutic effects
- Detect modifications of prespecified primary outcome measures
- Coordination of research trials and research groups
- Reduce idiosyncrasy in research and in IRB decision-making
- Provide orientation for IRB review
- Compensate for the shortcomings of early psychosurgery
- Attract industry funders

Source: Adopted from (Synofzik et al. 2012)

(Zarin and Tse 2008). If the DBS registry entails also post-intervention observational information covering periods of time longer than typically studied in controlled trials, it will provide an invaluable resource for longitudinal assessment of both efficacy and safety and the late unanticipated adverse event, as, for example, in patients with depression, where long-term data are currently still missing and often difficult to assess (Dunner et al. 2006). Such aggregation of information could provide benefits to trial participants who have agreed to participate in investigative DBS and who have placed themselves at risk not only due to the hope of personal benefit, but also due to the desire to increase generalizable medical knowledge (DeAngelis et al. 2005; Schlaepfer and Fins 2010).

A DBS registry could also serve as a repository for information that may not have ever been published but which might remain useful to the broader investigative community. For example, it would allow for the collection of ineffective DBS and help avoid a success bias which results from publication of only the efficacious studies. Knowledge of the failures as well the benefits will help avoid subject exposure to repeat studies that have already proven to be disproportionate and avoid the waste associated with repeating ineffective studies.

A DBS registry also allows to identify and aggregate information on DBS side effects which turn out to be therapeutic effects for comorbid features (e.g., remission of concomitant alcohol dependence in a case of nucleus accumbens DBS which primarily aimed at treating an anxiety disorder(Kuhn et al. 2007)). Moreover, it allows to detect modifications of prespecified primary outcome measures. Sometimes one study's adverse event becomes another's primary outcome (Schlaepfer and Fins 2010). For example, a case of unsuccessful DBS treatment of obesity was primarily published because of enhancement of memory states (Hamani et al. 2008). This point to the possibility that primary outcome measures might potentially be modified in the course of those DBS studies where the initial primary outcome measures are not achieved. If the primary prespecified outcome measures are posted in a DBS registry already at trial inception, such modifications can easily be detected.

Finally, a DBS registry could also serve an important ethical and regulatory role. It could help facilitate coordination of research trials by providing cross-center knowledge about past and ongoing studies and exert a moral corrective function through scientific consensus and common standards. By this means, it acts against possible idiosyncratic intervention and assessment standards and aberrant patient selection in single DBS centers across the world. By ensuring that all scientific centers adhere to a certain set of minimal standards, within and across center, variance will be decreased and less patients will receive unjustified harm.

Conclusion and Future Directions

It is primarily not a *scientific*, but an *ethical* necessity to ensure more rigorous methodological quality of DBS studies, comprehensive standards of DBS outcome reporting, prevention of selective reporting, and safeguards constraining a premature spread of DBS to novel applications and less experienced centers.

This is of particular importance with respect to those novel DBS applications that include children and/or complex neurodegenerative and neuropsychiatric diseases with highly vulnerable patients. Here we argued for a DBS research ethics program. As a first step, we illustrated the ethically driven need for optimal trial designs, common standards of outcome reporting, and registration also of single-case and small case studies. Future work has to work out these proposals in more detail and ensure and test their implementation in policy making, research, and clinical practice. So far, none of them have been put into practice.

Cross-References

▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Ethics of Sham Surgery in Clinical Trials for Neurologic Disease
▶ Human Brain Research and Ethics

References

Agid, Y., Schupbach, M., Gargiulo, M., Mallet, L., Houeto, J. L., Behar, C., & Welter, M. L. (2006). Neurosurgery in Parkinson's disease: The doctor is happy, the patient less so? *Journal of Neural Transmission. Supplementum, 70*, 409–414.

Bell, E., Mathieu, G., & Racine, E. (2009). Preparing the ethical future of deep brain stimulation. *Surgical Neurology, 72*(6), 577–586.

Bewernick, B. H., Hurlemann, R., Matusch, A., Kayser, S., Grubert, C., Hadrysiewicz, B., & Schlaepfer, T. E. (2010). Nucleus accumbens deep brain stimulation decreases ratings of depression and anxiety in treatment-resistant depression. *Biological Psychiatry, 67*(2), 110–116.

Castelnau, P., Cif, L., Valente, E. M., Vayssiere, N., Hemm, S., Gannau, A., . . . Coubes, P. (2005). Pallidal stimulation improves pantothenate kinase-associated neurodegeneration. *Annals of Neurology, 57*(5), 738–741. doi:10.1002/ana.20457.

Clausen, J. (2009). Man, machine and in between. *Nature, 457*(7233), 1080–1081.

Clausen, J. (2010). Ethical brain stimulation – neuroethics of deep brain stimulation in research and clinical practice. *The European Journal of Neuroscience, 32*, 1152–1162.

Cornu, C., Kassai, B., Fisch, R., Chiron, C., Alberti, C., Guerrini, R., & Nabbout, R. (2013). Experimental designs for small randomised clinical trials: An algorithm for choice. *Orphanet Journal of Rare Diseases, 8*, 48. doi:10.1186/1750-1172-8-48.

Coultas, D. (2007). Ethical considerations in the interpretation and communication of clinical trial results. *Proceedings of the American Thoracic Society, 4*(2), 194–198; discussion 198–199.

de Zwaan, M., & Schlaepfer, T. E. (2013). Not too much reason for excitement: Deep Brain Stimulation for Anorexia Nervosa. *European Eating Disorders Review*. doi:10.1002/erv.2258.

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., & Van Der Weyden, M. B. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *JAMA: The Journal of the American Medical Association, 292*(11), 1363–1364.

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., & Van Der Weyden, M. B. (2005). Is this clinical trial fully registered? A statement from the International

Committee of Medical Journal Editors. *JAMA: The Journal of the American Medical Association, 293*(23), 2927–2929.
Dickersin, K., & Rennie, D. (2003). Registering clinical trials. *JAMA: The Journal of the American Medical Association, 290*(4), 516–523.
Dunner, D. L., Rush, A. J., Russell, J. M., Burke, M., Woodard, S., Wingard, P., & Allen, J. (2006). Prospective, long-term, multicenter study of the naturalistic outcomes of patients with treatment-resistant depression. *The Journal of Clinical Psychiatry, 67*(5), 688–695.
Fins, J. J. (2000). A proposed ethical framework for interventional cognitive neuroscience: A consideration of deep brain stimulation in impaired consciousness. *Neurological Research, 22*(3), 273–278.
Franzini, A., Marras, C., Ferroli, P., Bugiani, O., & Broggi, G. (2005). Stimulation of the posterior hypothalamus for medically intractable impulsive and violent behavior. *Stereotactic and Functional Neurosurgery, 83*(2–3), 63–66.
Hamani, C., McAndrews, M. P., Cohn, M., Oh, M., Zumsteg, D., Shapiro, C. M., & Lozano, A. M. (2008). Memory enhancement induced by hypothalamic/fornix deep brain stimulation. *Annals of Neurology, 63*(1), 119–123.
Hariz, M., Blomstedt, P., & Zrinzo, L. (2013). Future of brain stimulation: New targets, new indications, new technology. *Movement Disorders*. doi:10.1002/mds.25665.
Hutton, P. (2013). Deep brain stimulation for anorexia nervosa. *Lancet, 328*, 305–306.
Kuhn, J., Lenartz, D., Huff, W., Lee, S., Koulousakis, A., Klosterkoetter, J., & Sturm, V. (2007). Remission of alcohol dependency following deep brain stimulation of the nucleus accumbens: Valuable therapeutic implications? *Journal of Neurology, Neurosurgery, and Psychiatry, 78*(10), 1152–1153.
Kuhn, J., Bauer, R., Pohl, S., Lenartz, D., Huff, W., Kim, E. H., & Sturm, V. (2009). Observations on unaided smoking cessation after deep brain stimulation of the nucleus accumbens. *European Addiction Research, 15*(4), 196–201.
Kupsch, A., Benecke, R., Muller, J., Trottenberg, T., Schneider, G. H., Poewe, W., . . . Volkmann, J. (2006). Pallidal deep-brain stimulation in primary generalized or segmental dystonia. *The New England Journal of Medicine, 355*(19), 1978–1990.
Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Zumsteg, D., Wennberg, R., Keren, R., . . . Lozano, A. M. (2010). A phase I trial of deep brain stimulation of memory circuits in Alzheimer's disease. *Annals of Neurology, 68*(4), 521–534.
Li, P., Huang, R., Song, W., Ji, J., Burgunder, J. M., Wang, X., . . . Shang, H. F. (2012). Deep brain stimulation of the globus pallidus internal improves symptoms of chorea-acanthocytosis. *Neurological Sciences, 33*(2), 269–274. doi:10.1007/s10072-011-0741-y.
Lipsman, N., Woodside, D. B., Giacobbe, P., Hamani, C., Carter, J. C., Norwood, S. J., . . . Lozano, A. M. (2013). Subcallosal cingulate deep brain stimulation for treatment-refractory anorexia nervosa: A phase 1 pilot trial. *Lancet, 381*(9875), 1361–1370. doi:10.1016/S0140-6736(12)62188-6.
Mallet, L., Polosan, M., Jaafari, N., Baup, N., Welter, M. L., Fontaine, D., . . . Pelissolo, A. (2008). Subthalamic nucleus stimulation in severe obsessive-compulsive disorder. *The New England Journal of Medicine, 359*(20), 2121–2134.
Mammis, A., Pourfar, M., Feigin, A., & Mogilner, A. Y. (2012). Deep brain stimulation for the treatment of tremor and ataxia associated with abetalipoproteinemia. *Tremor and Other Hyperkinetic Movements (New York, N.Y.), 2*.
Mayberg, H. S., Lozano, A. M., Voon, V., McNeely, H. E., Seminowicz, D., Hamani, C., & Kennedy, S. H. (2005). Deep brain stimulation for treatment-resistant depression. *Neuron, 45*(5), 651–660.
Moher, D., Schulz, K. F., & Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA: The Journal of the American Medical Association, 285*(15), 1987–1991.
Muller, U. J., Voges, J., Steiner, J., Galazky, I., Heinze, H. J., Moller, M., & Kuhn, J. (2013). Deep brain stimulation of the nucleus accumbens for the treatment of addiction.

Annals of the New York Academy of Sciences, 1282, 119–128. doi:10.1111/j.1749-6632.2012.06834.x.

Okun, M. S., Mann, G., Foote, K. D., Shapira, N. A., Bowers, D., Springer, U., . . . Goodman, W. K. (2007). Deep brain stimulation in the internal capsule and nucleus accumbens region: Responses observed during active and sham programming. *Journal of Neurology, Neurosurgery, and Psychiatry, 78*(3), 310–314.

Schiff, N. D., Giacino, J. T., Kalmar, K., Victor, J. D., Baker, K., Gerber, M., . . . Rezai, A. R. (2007). Behavioural improvements with thalamic stimulation after severe traumatic brain injury. *Nature, 448*(7153), 600–603.

Schlaepfer, T. E., & Fins, J. J. (2010). Deep brain stimulation and the neuroethics of responsible publishing: When one is not enough. *JAMA: The Journal of the American Medical Association, 303*(8), 775–776.

Schlaepfer, T. E., Cohen, M. X., Frick, C., Kosel, M., Brodesser, D., Axmacher, N., . . . Sturm, V. (2008). Deep brain stimulation to reward circuitry alleviates anhedonia in refractory major depression. *Neuropsychopharmacology, 33*(2), 368–377.

Schupbach, M., Gargiulo, M., Welter, M. L., Mallet, L., Behar, C., Houeto, J. L., . . . Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*(12), 1811–1816.

Synofzik, M., & Schlaepfer, T. E. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*(12), 1511–1520.

Synofzik, M., & Schlaepfer, T. E. (2010). Neuromodulation – ECT, rTMS, DBS. In H. Helmchen & N. Sartorius (Eds.), *Ethics in psychiatry. European contributions* (pp. 299–320). Heidelberg: Springer.

Synofzik, M., & Schlaepfer, T. E. (2011). Electrodes in the brain-ethical criteria for research and treatment with deep brain stimulation for neuropsychiatric disorders. *Brain Stimulation, 4*(1), 7–16.

Synofzik, M., Fins, J. J., & Schlaepfer, T. E. (2012). A neuromodulation experience registry for deep brain stimulation studies in psychiatric research: Rationale and recommendations for implementation. *Brain Stimulation, 5*(4), 653–655. doi:10.1016/j.brs.2011.10.003.

Warnke, P. (2013). Deep brain stimulation for movement disorders: What counts in the end is the end result. *Journal of Neurology, Neurosurgery, and Psychiatry*. doi:10.1136/jnnp-2013-305832.

Wu, H., Van Dyck-Lippens, P. J., Santegoeds, R., van Kuyck, K., Gabriels, L., Lin, G., . . . Nuttin, B. (2013). Deep-brain stimulation for anorexia nervosa. *World Neurosurgery, 80*(3–4), S29 e21-S29 e10. doi:10.1016/j.wneu.2012.06.039.

Zarin, D. A., & Tse, T. (2008). Medicine. Moving toward transparency of clinical trials. *Science, 319*(5868), 1340–1342.

39 Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review

Anna Pacholczyk

Contents

A. Pacholczyk
Centre for Social Ethics and Policy, School of Law, University of Manchester, Manchester, UK

Institute for Science Ethics and Innovation, Faculty of Life Sciences, University of Manchester, Manchester, UK
e-mail: anna.pacholczyk@postgrad.manchester.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_29,

Abstract

Over the last decades, deep brain stimulation has become an established treatment for movement disorder. However, the recent clinical research investigating the application of the technique for neuropsychiatric disorders as well as the potential to use DBS in a wider range of conditions has brought ethical controversy. This chapter examines some of the ethical objections against extending the use of DBS beyond the treatment of movement disorders. The first part of this chapter focuses on the objections related to the past uses of psychosurgery and examines the ethical weight of invasiveness and reversibility, two commonly used concepts in ethical discussions about DBS. The second part examines objections related to freedom and autonomy, including the slippery slope arguments related to "mind control," as well as the issue of control over the parameters of the stimulation in the context of cognitive liberty. The third part briefly discusses the objection to the enhancement uses of DBS, highlighting different contexts in which DBS can be used for enhancement while arguing for a systematic consideration of ethically relevant factors instead of the reliance on the treatment/enhancement distinction. The final part addresses the objection to DBS related to threats to identity. First, the ethical weight of discontinuity under psychological continuity approaches to identity is discussed. Next, the focus turns to the narrative approach to identity and Schechtman's objection to the use of DBS, and it is argued that although the narrative approach brings valuable insights, the normative conclusions we can draw from it are limited.

Introduction

Although deep brain stimulation (DBS) has been used to treat Parkinson's disease for some time (see ▶ Chap. 35, "Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects"; ▶ Chap. 63, "Ethics of Functional Neurosurgery"; ▶ Chap. 29, "Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives"), recent decades have seen an increased focus on the potential of DBS as treatment for neuropsychological disorders such as major depression, OCD, and Tourette's syndrome (Holtzheimer et al. 2012; Krack et al. 2010). Preliminary research investigated the use of DBS for disorders of consciousness (Schiff et al. 2007; Yamamoto et al. 2010), Alzheimer's disease (Laxton et al. 2010; Smith et al. 2012; Lipsman et al. 2013), obesity (Halpern et al. 2008; Bauer et al. 2008; Kuhn et al. 2007; Lu et al. 2009), aggression (Hernando et al. 2008; Kuhn et al. 2008), and a handful of case studies reported post-DBS outcomes in neuroacanthocytosis and Huntington's disease (Edwards et al. 2012).

Some estimate that over 80,000 patients have already been treated with DBS and another 8,000–10,000 undergo stimulation each year (Lozano 2012; see ▶ Chap. 34, "Ethical Implications of Brain Stimulation"). The investigation of DBS for neuropsychological and neurological conditions is likely to increase. Yet, the prospect of extending the use of DBS to the treatment of depression, OCD, Tourette's

syndrome, and addiction, as well as the potential for enhancing interventions has led to significant unease and criticisms. This chapter critically reviews some objections to the use of DBS related to research ethics (the past abuses of psychosurgery, invasiveness), freedom and autonomy, as well as personal identity (psychological continuity and narrative approaches).

Psychosurgery, Invasiveness, and Reversibility

The Shadow of Psychosurgery: "DBS Should Not Be Used for Neuropsychiatric Disorders Because of Past Abuses in the Use of Psychosurgery and ECT in That Population"

DBS has developed in the shadow of past psychosurgery abuses, including the lobotomies performed by Moniz and Freeman and Heath's Tulane electrical stimulation program (Baumeister 2000; Pressman 1998). Moreover, dystopian images from works of popular fiction picturing brain interventions, such as *One Flew Over the Cuckoo's Nest* or *The Manchurian Candidate*, may influence the public perception of DBS (Schermer 2011). Although the current media reporting is mainly positive ("pacemaker for the brain" rather than "Brave New World," see Racine et al. 2007), there is a worry that DBS treatments will mirror the history of psychosurgery, with the initial hype leading to misuse and overuse subsequently replaced by controversy and abandonment of the technique (Hariz and Hariz 2012). According to Hariz and Hariz (2012), suggesting "that the technique be used for cognitive enhancement in non-pathological conditions, and even for alleged antisocial behaviour, are reminiscent of a dark era, and constitute a bad omen for the future of this technique."(p. e5477)

Many commentators acknowledge the history of psychosurgery but have distinguished the current incarnation of DBS from past uses, especially the misused application of lobotomy (e.g., Bell and Racine 2012). DBS is significantly more reversible, its settings are adjustable, the knowledge base supporting it is more solid, and it is performed in the context of different ethical and professional standards of practice, including the standards for patient selection, informed consent, and ethical oversight (for a systematic comparison, see Table 1 in Synofzik and Schlaepfer 2008). Others criticize the "clear-cut" approach to the history of DBS and psychosurgery, reinforced by some prevalent misconceptions (Hariz et al. 2010). While the "clear-cut with the past" narrative might protect DBS from unjust controversy and shallow comparisons, it might rob us of valuable lessons and come at the price of misrepresenting facts. On the other hand, while past questionable practices might give us a good reason to proceed with caution and carefully monitor DBS applications, they are not in themselves a good reason to halt the clinical research on novel applications.

Moreover, solely focusing on the history of psychosurgery in discussing the ethics of DBS brings three dangers. Firstly, shallow comparisons based on the invasiveness and past inappropriate use might trivialize the debate, while it is

necessary to carefully examine the reasons for the past misuses of psychosurgery, asking which aspects of those practices were and were not ethically problematic in order to see to what extent we have progressed since then (see ▶ Chap. 59, “Ethics in Neurosurgery”; ▶ Chap. 31, “Informed Consent and the History of Modern Neurosurgery”; ▶ Chap. 60, “Neurosurgery: Past, Present, and Future”; ▶ Chap. 36, “Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia”). Secondly, we may overlook a host of different ethical challenges we face today, such as those related to distribution of scarce resources and proceeding with clinical research in a market economy with the strong and necessary involvement of the medical devices industry from the earliest stages of application. Thirdly, we might miss some more important and current comparisons, such as the more recent history of psychopharmacological agents and the underreporting of long-term side effects (see ▶ Chap. 38, “Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations”).

Intuitions About Invasiveness: “DBS Should Not Be Used for Neuropsychological Disorders Because It Is Invasive”

Reversibility and invasiveness of DBS are the two key words repeatedly used in both scientific and ethical papers about DBS. The gist of the use can be captured in one sentence: “Despite the fact that DBS is invasive (which is obviously bad), it is reversible (which is very good).” Although the two terms are commonly used as if their meaning was obvious, the root of their rhetorical power is somewhat more complex. Arguments against the use of DBS and calls for caution often include references to the invasiveness of the procedure. For example, Lipsman and colleagues (2011) pointed out that “[a]lthough DBS is [an] minimally invasive neurosurgery, it is the most invasive psychiatric treatment available” (p. 40). But what is the normative force of the concept and what could “invasiveness” refer to?

Firstly, “invasiveness” of the procedure might refer to a particular mode of crossing the skin boundary, as in the Merriam-Webster Medical Dictionary which defines “invasive” as one “involving entry into the living body (as by incision or by insertion of an instrument)” (“Invasive,” n.d.). Clearly, crossing the skin boundary is not in itself a good or a bad thing and a normatively rich concept of invasiveness refers to certain kinds of harm. In DBS, invasiveness points mainly to surgery- and device-related adverse events, including the risk of intracerabral hemorrhage, postoperative infection, and electrode displacement (Kuhn et al. 2010). Although invasiveness might point toward risks associated with the surgery, there are certainly procedures that are invasive in a technical-medical use of the word, and in ordinary circumstances not terribly dangerous (such as blood draw or biopsies). Thus, although invasiveness might point toward risk, it is not a clear normative guide.

The rhetorical force of the word “invasiveness” in the context of DBS is also rooted in several other sources. Firstly, it evokes intuitions about bodily boundaries,

such as those echoed in the emphasis put on concepts such as bodily integrity in the context of cadaver organ donation and a certain intuitive bias against surgical procedures. On the other hand, invasiveness might be a psychological concept and refer to the effects of the intervention on cognitive and affective processes and corresponding physiology. For example, psychotherapy might be considered to be noninvasive in one sense (does not involve incisions, etc.), yet invasive in the sense of changing thought and emotional processes, with associated physiological changes such as changes in brain metabolism. Moreover, invasiveness may refer to inappropriateness of a certain external influence – questions that cross the boundary of privacy might be called invasive and psychotherapy that does not take into account patients' views and wishes might be called invasive.

Neuromodulation techniques are sometimes, explicitly or implicitly, evaluated on the basis of invasiveness: the more invasive the neuromodulation technique, the less acceptable its use. Disguised as a straight-forward descriptive term referring to a kind of medical intervention but commonly used with strong normative undertones, the notion of invasiveness is a poor normative guide. Since the concept's rhetorical force comes from several separate sources beyond simply harms associated with surgery, it is important to be aware of how it is used.

Dilemmas in Patient Selection: "DBS Should Only Be Used as a Last-Resort Treatment"

Kuhn et al. (2009) note that there is a deep intuition, reflected in ethical guidelines, that an invasive procedure such as DBS should only be a last-resort treatment. Whether this widely stated conviction is justified is far from evident. It is one thing to emphasize the potential of DBS for helping patients with severe and otherwise irremediable disorder and prioritize early-stage clinical research to target this patient population on the basis of both need and risk/benefit ratio. It is a different, and a much stronger claim, to state that the procedure should *only* be tried when nothing else worked.

Since the emergence of the "access to experimental treatment" movement and an increasing consideration for patient treatment preferences and the recognition that patient's toleration of different kinds of risks differs, the protection of patients has ceased to be the only argument on the table. It is conceivable that in the not-so-far future, some subpopulation of patients might prefer to resort to DBS before trying all, many, several, or any of the available courses of different plausible pharmacological remedies (see ▶ Chap. 37, "Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine"). We might very soon be forced to decide whether to provide access to patients with a less impressive *résumé* of tried treatments. DBS plus drugs has already been suggested to be a better treatment option than the current pharmacological treatment for early motor complications in Parkinson's disease: Despite a higher rate of adverse events, the neurostimulation group had a higher mean score on a quality of life measure (Schuepbach et al. 2013). References to invasiveness and risk usually do not tell

the whole story and if we deny access, it is important to be clear about the reasons. Is it a matter of prioritizing the population with the greatest severity of symptoms? In the light of the troubled past of brain stimulation and surgery, is it a matter of protecting DBS from controversy to protect less controversial applications but at a cost of denying other patients access to a potential benefit? Is it prioritizing research on the population that is likely to achieve greatest symptom reduction and subsequently demonstrate the greatest efficacy of treatment?

Although it is sometimes easier to reach consensus about ethical guidelines than on the reasons for those guidelines, it is important that the reasoning behind the guidelines is made explicit. The strength of the rationale for those guidelines changes as the research progresses, and, as the consensus starts to dissolve, the transparency about reasons is necessary for the quality of our ethical discussion.

Intuitions About Reversibility

"Reversibility" is an often evoked concept, typically used to emphasize the comparative advantage of DBS over ablative surgery (e.g., Andrade et al. 2010; Greenberg et al. 2006). The intuition about the prudential and moral force of reversibility might be rooted in the intuition that ceasing an active intervention alleviates the associated harms. Prudentially, a risky intervention is seen as less risky if we can "undo" the effects. Ethically, one could hold a view that to reverse the effects of an intervention is to "wipe out" some of the moral implications of it: One is less morally culpable if the harm that was done is averted.

One of the concepts of reversibility refers to return to the *status quo ante* – back to the starting point. Yet, such understood reversibility does not wipe out all moral implications: Even restoring the *status quo ante* does not alleviate the moral weight of the harms (or benefits) already incurred.

Moreover, a comparatively lesser destruction of brain tissue does not necessarily equal nearly full reversibility of morally relevant effects. The reversibility of the intervention itself is not the same as the reversibility of the *effects* of the intervention. Firstly, the physiological and functional effects directly traceable to the effects of stimulation might be reversible only to some extent. The reports of a rapid increase in treated symptoms after the stimulation was stopped (e.g., after battery failure) could lead us to assume that all other effects are also reversible. However, this conclusion might be premature, especially in the absence of solid empirical data about long-term effects after discontinuation of treatment as well as the precedence of initially under-reported and overlooked side effects of pharmacotherapy (e.g., SSRIs), sometimes remaining for months or years after the treatment has been discontinued (Csoka and Shipko 2006).

Secondly, a patient's psychosocial situation might have changed. The patient might have spent a significant time adapting to a new state of being, developed new preferences and ways of living, lost the habits and environmental support that allowed them to cope previously, and developed a new set of strategies to compensate for a different profile of functional strengths and deficits. Thus, even if we

could know that otherwise the patient had returned to the exact same affective and cognitive state as she was in prior to DBS, the patient might be left in an even less advantageous position than prior to DBS. Naturally, the inverse is also possible – a patient might be in a generally better position to cope with the consequences of discontinuing treatment. The case in point is that the removal of the DBS electrodes does not entail returning to the starting point – the procedure is, in Synofzik and Schlaepfer (2011) words, "psychosocially irreversible" (p. 10).

While *the stimulation* might be reversible, it remains an open question to what extent *the effects* of the stimulation are indeed reversible, and it is the effects which are morally relevant. It might be better to regard DBS to be as reversible as marriage is – although one can *divorce*, it would be an exaggeration to say that marriage is *reversible*.

Freedom and Autonomy

"Deep Brain Stimulation Should Not Be used Because It Can Lead to Mind Control"

Extending the use of DBS can lead to worry about a slippery slope leading to "mind control." Although the dystopian image of externally controlled soldiers or prisoners forced to undergo DBS for behavioral control is indeed troubling, several intuitions seem to feed into the "mind control" worry. Firstly, one can object to the coercive use of DBS (see also ▶ Chap. 57, "Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent"). Secondly, the worry might refer to the use of medical technologies for social control purposes (see ▶ Chap. 79, "The Morality of Moral Neuroenhancement"). Thirdly, we might see a difference between the use of coercion or a strong incentive aimed at a change in the sphere of the physical (e.g., interrogation techniques aimed at signing a declaration of agreement with the ideology X) and the mental (techniques aimed at brainwashing prisoners into believing in the tenets of ideology X). Physical coercion does not preclude retaining autonomy over one's thoughts and disobedience when an opportunity arises. As Mahatma Gandhi is thought to say: "You can chain me, you can torture me, you can even destroy this body, but you will never imprison my mind."

Further, we might perceive a difference between an indirect impact on the brain (manipulating one into believing X) and a direct one (using drugs or brain stimulation together with propaganda material to induce believing X). The perception here may be that while indirect impact can be modulated by beliefs, direct impact leaves less ability to modify the behavioral consequences. This is why, although DBS is subject to a similar dual-use problem as knifes and police force, the prospect of misuse of direct means of brain modulation and modification evokes especially strong unease.

The "mind control" objection relies on a slippery slope argument. Thus, the strength of the objection depends on what particular technologies are likely to do as well as practical matters such as comparative cost. For example, the prospect of a remotely steered individual performing complex actions is significantly more

far-fetched than using DBS for evoking undirected aggressive behavior or for remote disabling of movement. Even if the technology could be developed, it might be financially impractical to use DBS instead of stun guns and existing surveillance. Secondly, the strength of this objection depends on the balance and probability of harms and benefits, a factor perennially tricky to asses in dual-use dilemmas (see ▶ Chap. 116, "Biosecurity as a Normative Challenge"; ▶ Chap. 112, "Weaponization of Neuroscience"). The distribution of harms should also be considered. Coercive use of DBS for social control purposes is less likely to happen on any significant scale and without controversy in most Western democracies, but the situation may look different elsewhere. Thirdly, there is a question of whether halting the research and treatment would make a difference in the longer run in reducing the risk of misuse.

"Deep Brain Stimulation Should Not Be Used Because It Might Compromise Freedom"

Moreover, there is a wider issue of liberty with respect to one's thoughts and desires. Even if the overall consequences of DBS applied sparingly and mainly to a good social purpose, as Persson and Savulescu (2008) propose, we might not want to live in a society that overwhelmingly trades freedom for safety (Harris 2011).

Actions are subject to legal and moral prohibition, not desires and thoughts on their own. The freedom to do as we please in the domain of our mental life may be of particular importance (see ▶ Chap. 42, "Mind Reading, Lie Detection, and Privacy"). This is an underlying thought in Sententia's (2004) formulation of cognitive liberty:

> "*Cognitive liberty is a term that updates notions of 'freedom of thought' for the 21st century by taking into account the power we now have ... to monitor and manipulate cognitive function. Cognitive liberty is every person's fundamental right to think independently, to use the full spectrum of his or her mind, and to have autonomy over his or her own brain chemistry*" (pp. 222–223).

While the coercive use of DBS may compromise liberty, voluntary use of DBS might be regarded as simply exercising that liberty. It is a matter of discussion whether cognitive liberty is merely a negative freedom (freedom from interference) or a positive freedom (freedom to) and how strong an argument it provides for the access to relevant technologies. Moreover, there is a question of justifiable restrictions to cognitive liberty (see ▶ Chap. 32, "Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry"). Currently, this question arises in relation to access as well as the choice of stimulation settings.

Who Chooses the Settings?

The first question is whether one should have control over the stimulation and be able to adjust the settings out of the consulting room. Generally, patients can control

whether the device is on or off. Some choose whether the device is on or off depending on an activity (e.g., walking: on, talking: off), and some can also turn the device down during the night to save battery life (Mathews et al. 2011). Patients whose treatment influences mood might want to modulate their mood appropriately to the situation at hand. If DBS influences mood in more complex ways than just changing the baseline of positive/negative affect, control over settings might be necessary to experience specific valued moods. At the moment, there is no data on whether leaving the control of the stimulator up to the patient is beneficial or harmful (Mathews et al. 2011), although some have worried about excessive stimulation for hedonic purposes (Oshima and Katayama 2010). Regardless of harms and benefits, continuous control over the stimulation could be seen as an important way of exercising autonomy.

Secondly, during the adjustment of stimulation settings, patients and doctors may face a dilemma as to what level of mood improvement is optimal. What should happen if their opinions differ? The importance of a careful consideration of those questions is demonstrated in a recently reported case of a patient receiving NA (Nucleus Accumbens) stimulation for OCD and GAD. As the voltage increased, the sensation of being relaxed and hedonic experience increased, while at higher voltages, the patient began to feel "unrealistically good," "overwhelmed," and began to worry that the anxiety was going to return (Synofzik et al. 2012). Although initially an intermediate voltage for stimulation was chosen, the next day, the patient requested adjustment of the settings so that he could feel "a bit happier."

Especially when the treatment is received for anxiety disorder or depression, it might be difficult to delineate what should count as treatment and what as enhancement; moreover, it is unclear why we should care about that distinction at all. We should instead, as Synofzik et al. (2012) have argued, focus on what stimulation parameters are optimal in promoting patients' general well-being. Moreover, each mentally competent person should at least in principle be free to decide and act according to his or her personal preferences; arguments need to be given in the first place not for allowing the use of DBS enhancement, but for prohibiting it (Pacholczyk 2011; Synofzik et al. 2012). Further questions include: To what extent, emotion or mood enhancement promotes long-term well-being, psychosocial adjustment, functioning in the society, and happiness (see ▶ Chap. 76, "Ethics of Pharmacological Mood Enhancement"); which of those outcomes should be prioritized when they diverge; and who is to decide which of those outcomes should have priority.

Nontherapeutic Use of DBS: "DBS Should Not Be Used for Enhancement"

The case reported in Synofzik et al. (2012) brings to the fore the question about the permissibility of extending the use of DBS to enhancement uses. The "treatment of any disorder, including a mental disorder, is ethically permissible" argument legitimizes the use of DBS for accepted mental health disorders, excludes non-treatment uses, and leaves us uncertain about what to do in cases when medicalization of an

issue is especially strongly challenged (e.g., overeating or addictions). However, it relies on a problematic treatment/enhancement distinction (see ▶ Chap. 75, "Neuroenhancement"; ▶ Chap. 80, "Reflections on Neuroenhancement"). If we see the treatment/enhancement distinction as *normatively* relevant only to the extent that it traces cost-benefit ratio, the argument above could be simplified and strengthened by omitting the "treatment" criterion and instead simply focusing on the benefit (Synofzik and Schlaepfer 2008). Focusing on the benefit to a particular individual means that enhancement uses cannot be easily discounted ethically and at the very least warrant a careful consideration (Pacholczyk 2011).

As several authors put forward, it is ethically problematic to pursue brain interventions for frivolous reasons and this especially applies to relatively high-risk procedures like DBS (e.g., DeGrazia 2005a; Focquaert and DeRidder 2009). Since not many would encourage acting on frivolous reasons, there is a question of why there is an implied association between nontherapeutic uses and frivolity and who exactly decides what is and is not a frivolous reason.

For the purpose of the argument, let us consider three kinds of cases where a nontherapeutic use would be permissible on the basis of the cost-benefit ratio. Firstly, some patients who have already undergone DBS for treatment purposes may wish to control the stimulation settings for enhancement. The surgery risks have been already been taken, and so, the cost-benefit analysis will only include the risks associated with flexible stimulation played against possible benefits. Secondly, in some (admittedly rare) cases, a person might stand to greatly benefit from DBS if successful, and, while understanding the high risks, be willing to take them.[1] As argued by Pacholczyk (Pacholczyk 2011), the individual does not have the obligation to convince others that their values, acceptance of risks, and value attached to certain benefits are correct – what they are required to do is to know, understand, and weigh the risks, uncertainty, and the kinds of benefits that are likely. As long as they do, it is they who determine whether an intervention is worth trying. Thirdly, a nontherapeutic (but also technically non-enhancing) use might be envisaged when individuals want to be implanted with DBS for reasons having to do with pushing the boundaries of science and medicine. For example, a neuroscientist might want to perform a variety of cognitive tests, while she undergoes DBS to generate new knowledge. Then DBS would not be in that person's medical interest, and perhaps not in that person's narrowly construed self-interest (as there may be no strong expectation of improvement in cognitive or emotional capacities that would benefit the individual), but still can be in a widely understood self-interest that reflects their values related to the exploration of scientific questions. The three discussed cases demonstrate that it is unjustified to discount all enhancement uses as "frivolous" or necessarily ethically problematic on the basis of cost-and-benefit analysis alone.

However, DBS neurosurgeons face another set of questions before endorsing nontherapeutic use. Some questions and issues to consider include: (a) the fact that

[1]Doctors' and society's reasons to either provide access to the intervention or not are a separate issue

refusing DBS might in fact mean *denying benefit*, and therefore should be carefully considered and justified (or re-considered), (b) questions about who funds the physician's time and costs associated with surgery and the follow-up, (c) issues of priority: What place should be given to nontherapeutic uses in the context of limited expertise and physician time; this question fits into a wider array of issues that has to do with prioritizing therapy and clinical research with patients from different patient groups, suffering from conditions of varying severity and urgency, (d) the influence that nontherapeutic use might have on the public attitudes toward DBS in so far as they influence the development and application of the technique.

Identity Objections to DBS

Since the use of DBS in Parkinson's disease, an increasing number of commentators have raised and explored philosophical and ethical worries about its impact on selves, personalities, and persons (Hildt 2006; Glannon 2009; Müller and Christen 2011; Lipsman et al. 2009; Schechtman 2010; Synofzik and Schlaepfer 2008; see ► Chap. 25, "Impact of Brain Interventions on Personal Identity"; ► Chap. 22, "Neuroethics and Identity"; ► Chap. 23, "Neurotechnologies, Personal Identity, and the Ethics of Authenticity"). Merkel et al. (2007) point to the widespread philosophical worry that one's personal identity might be compromised as a result of brain interventions: "The fear is often expressed that an individual may no longer be "the same person" he or she used to be prior to an intervention in the brain" (p. 4). The concern about the altering of fundamental aspects of one's self, and thus one's personal identity, might be especially salient in the case of DBS, because the changes can be effected relatively quickly and be clearly causally linked to DBS.

Although everyday conversations may casually include phrases such as "she wasn't herself," or "he is not the person he used to be," in philosophical discussions, "persons" and "selves" tend to be highly specific, technical concepts, which have been developed over decades, if not centuries, of intense examination of our intuitive responses to various cases and scenarios. The concept of numerical identity grounds an understanding of how something can continue to exist despite undergoing significant change (DeGrazia 2005b). However, the question of personal identity persistence is not only a question of numerical identity, but also concerns the identity of a particular *person*.

Psychological Continuity Approaches to Identity: "DBS for Neuropsychiatric Disorders Should Not Be Used Because it Threatens Personal Identity"

According to the psychological approach, our personal identity consists (at least partly) in some sort of psychological continuity. In most psychological theories (e.g., Locke 1975 [1694]; Parfit 1984; Perry 1972), the relevant type of continuity is a continuity of experiential contents, or, in other words, the maintaining of

psychological connections over time (Olson 2003). Examples of such psychological connections include having an experience and later remembering it, forming an intention and later acting on it, and the persistence of certain beliefs, desires, and character traits. In other psychological theories (e.g., Baker 2000; McMahan 2002; Unger 1990), by contrast, the relevant type of continuity is the continuity of one or more basic psychological capacities, which may persist despite loss of memories and other experiential contents (e.g., a basic capacity for reasoning or the capacity for conscious experience).

Yet, in either of these philosophical stances, it is unclear how disruptions in psychological continuity would translate into ethical objections or ethical reasons to act. To arrive at such an ethical imperative, we need further arguments: whether psychological continuity is valuable, and whether that value is sufficient to determine what we should do and why (Focquaert and DeRidder 2009; Synofzik and Schlaepfer 2008; Witt et al. 2011). Assume that for various reasons, life-shaping events, periods of reading too much of mind-tinkering philosophy, a marked change in social environment and social roles played, moving to a different country, an accident, illness, etc., the psychological continuity link is too weak to provide a sufficient connection between different selves. How that would translate into moral reasons to act remains unclear. Should I make sure I stay sufficiently similar to myself now? Should I change slowly to maintain sufficient psychological connections? Do I have an obligation to struggle to re-gain my previous way of being? Is an attempt to radically reinvent myself an unquestionably morally dubious enterprise? If I could radically reinvent myself and no other ethical objections apply, would it be wrong for me to reinvent myself solely in virtue of this being a change that threatens psychological continuity, or results in a different self? It is far from clear. Therefore, even if we assume that psychological connections are severed by DBS (and this is not clear), this does not necessarily imply an ethical imperative not to undergo DBS.

Perhaps, if we are going to drastically change career paths, have children, marry, permanently move countries, think about taking antidepressants or other medication or decide whether to undergo DBS, or make any other potentially personality- or identity-changing decision, we indeed should ask ourselves, "How is it going to change me? How do I understand myself? How comfortable am I with the prospect of radical change? Although reflecting on psychological approaches to personal identity gives an opportunity to reflect on how we think about *ourselves* and about *our selves*, it does not indicate that we should act in a certain way." Therefore, as with other significant decisions, the possibility of radical change is an *important consideration*, but a prima facie *weak objection* to the use of DBS specifically.

Narrative Identity: "DBS for Neuropsychiatric Disorders Should Not Be Used Because It Cannot Be Incorporated Into a Self-Constituting Narrative"

Narrative accounts of identity typically pose that in practice, we create the sense of who we are through the construction of an autobiographical self-narrative

(see ▶ Chap. 24, “Dissociative Identity Disorder and Narrative”). Those self-narratives are commonly thought to be means by which we order both voluntary and involuntary aspects of our experiences, give and re-evaluate their meaning and importance, and understand and give meaning to our lives. It is often posited that construing self-narratives is very important, if not necessary, for a full and flourishing life.

Schechtman’s Objection to the Use of DBS: “DBS Threatens Personal Narrative Identity Because Narratives After DBS Do Not Fulfill the Articulation Constraint”

An influential account of narrative identity, and one that has been explicitly applied to the ethical assessment of DBS, is that of Schechtman. Schechtman (1996) distinguishes a re-identification question (what makes someone the same person over time, as in Parfit 1984) and characterization question (what it is to be a particular person). In the narrative account of identity, the focus is on whether a subject “creates their identity by forming an autobiographical narrative—a story of his life” (1996, p. 113). Not all narratives can be identity-constituting and Schechtman proposes that for a narrative to be identity-constituting, it must satisfy two constraints. First is the articulation constraint (the person must be able to provide some account of her history, her life situation, and her motivations) and another is the reality constraint (the self-narrative must be coherent with basic facts about how the world is).

Schechtman (2010) considers the use of DBS to be a threat to narrative identity. Personality changes that result from DBS, Schechtman argues, are at odds with the articulation constraint on identity-constituting narratives according to which “the narrator should be able to explain why he does what he does, believes what he believes, and feels what he feels” (1996, p. 114). When the causes of actions can be traced back to the influence of DBS, patients will not be able to explain how their actions flow from their desires, beliefs plans, and projects – because they do not. Discussing a hypothetical case of a patient who experienced profound personality changes after DBS, she states that “his current passions and interests—the things he takes as reasons—were caused by manipulation of his brain” (2009, p. 85).

Arguments Against Schechtman’s Objection

However, Schechtman’s conclusion might be not satisfactory on her own account. Schechtman concludes that DBS threatens narrative identity despite acknowledging that “since narrative is a dynamic notion, continuity of narrative is thoroughly compatible with even quite radical change.” (2010, p. 140) As Baylis (2011) rightly notices, Schechtman’s account leaves open the possibility that personality changes can be successfully integrated into a subject’s autobiographical narrative.[2] There is no reason to think that a subject will necessarily be unable to provide a satisfactory

[2]This holds whether or not a change in a given trait was intended. Narratives are typically thought to incorporate both voluntary and involuntary aspects of experience.

account of her history, life situation, and motivations; to narrate parts of her life in a self-conscious way; to render her self-narrative intelligible. Suffering and fighting a disease, thinking about DBS as a treatment option, the process of consent to DBS, the period of adjustment of settings, etc., can all form part of such a narrative.

It therefore seems that Schechtman relies on an implicit assumption that the mechanism of change matters crucially; yet, the reader is left in the dark as to why the involvement of a technological or medical means should be viewed with a *special* suspicion. One can wonder why, if it coincidentally happened so that either a sudden and dramatic life experience or a month of practicing qigong produced exactly the same physiological changes in brain function (and resulted in the same profile of personality changes), we should view this "natural" way of personality change with lesser suspicion than the changes that result from DBS.

A more charitable reading of Schechtman may get at some of our intuitions as to why a change of values, beliefs, or character traits following DBS would be problematic: It is not that DBS is problematic in virtue of it being a technological means of affecting change, but rather that the intervention belongs to a class of change-affecting events which are *difficult to make sense of within a personal narrative*. Perhaps a shift in views after an intense qigong practice or a life-shaking event could be equally problematic, if unaccompanied by reflection and integration of the new stance toward life, including giving epistemic and genealogical reasons for this stance. However, it seems that Schechtman's claim – that changes after DBS *cannot* be incorporated into a narrative– is too strong. Perhaps, DBS can in some cases undermine a person's narrative, but if a person could reinvent a narrative following severe personality changes due to head trauma, there is no obvious reason that a person who undergoes DBS could not. Thus, rather than being a strong objection to DBS under all circumstances, Schechtman's highlights challenge that DBS could create.

Another related, but separate worry would be that of epistemic justification for the changed values, beliefs, and character traits. This epistemic problem could be especially important if new values, beliefs, and character were *less* justified than the previous ones. However, in situations when previous justifications were weak (e.g., a depressed patient could hold a belief about him or herself being worthless as a result of childhood trauma), justified in the past but not the present ("My life is full of emotional pain" for a formerly depressed patient) or equally well or more poorly justified than a new belief (e.g., "I'm a conservative because my father was a conservative," "I'm a liberal because that is my fancy after DBS"), the degree of justification of the beliefs does not change, which would further qualify the strength of the objection.

Arguments Against the Strong Ethical Narrative View and the Strong Psychological Narrative View

Moreover, one can criticize the ethical narrative premise which states that a coherent and continuous narrative is necessary (or at least necessarily highly conducive) to a flourishing life. For example, Vice argues that if we take the narrative view "seriously" and "literally," it requires that we cast ourselves as "characters—usually the protagonists—of the stories we tell or could tell about ourselves" (2003, p. 93) and

points out the harm of associated self-deception, the risk of constraining autonomy, and being inauthentic. Mackenzie and Poltera (2010) reject the conception of narrative self-constitution that underpins the "story-telling" critiques (Strawson 2004; Vice 2003) and argue that they present narrative accounts as more rigid and literal than they are. Perhaps Strawson's and Vice's arguments apply more readily to more demanding accounts of narrative identity.

For example, Ricoeur's (1984, 1985, 1988, 1992) account of narrative identity is sometimes seen as imposing stringent criteria of structural unity of the story and homophony, which, as critics argue, may mean that meaningful and important experiences would go unnoticed, trivialized, or repressed (e.g., as in Maan 2010; Muzak 2007). On the other hand, Ricoeur's approach can be seen as a response to a perceived postmodern fracturing of the subject, and loosening the criteria too much risks making the narrative indistinguishable from a description or a dream sequence, thereby resulting in the loss of what makes a narrative identity specifically *narrative*. This discussion rooted in the literary tradition has its analogue in discussions within the field of narrative psychotherapy (e.g., Hermans 2003; Hermans and Dimaggio 2004) and the issue of the stringency of criteria for a narrative account of identity remains open. Even if indeed the full blow of Vice's and Strawson's claims here are taken only by the more demanding narrative views, the questions about justifiability of the ethical narrative thesis and the psychological narrative thesis remain.

First, the question is whether people indeed think about themselves in narrative terms. Although the conclusion will depend on the identity criteria posited by a particular theory of narrative identity (a conceptual issue), it is also an empirical question. In the absence of convincing evidence either way, it is plausible to assume that there are some people who understand their own lives through an elaborate narrative and evoke it explicitly on regular basis, others who re-construe and evoke an otherwise transparent self-narrative only when prompted by circumstances, and those who do not construct their lives through and in a narrative that would fulfill criteria for an identity-narrative. The latter sort of people may not be particularly reflective; suffer from a mental health issue that undermines their capacity to form complex narratives; or be highly reflective but assign little importance to a consistent narrative line, while, perhaps, attending to current reflectively endorsed interests, desires, and preferences.

Although according to the ethical narrative view, the lives of the last group are *necessarily* less meaningful or flourishing, it is a very strong thesis. There may be an aesthetic allure in a Ricoeurian view of fully flourishing life as life that is "complete," but I doubt whether such an aesthetic preference translates into an epistemic or moral imperative to lead or even aim at such a life. Thus, the move from how *we might* construe ourselves through the process of creating an autobiographical identity narrative to the conclusion that construing of such personal narrative is *necessary* for a meaningful or flourishing life lacks strong justification. In medical practice, I see no strong reason to favor a narrative identity view (and accept the potential objections to DBS coming with it) *if* the patient does not endorse or care about it.

A related objection is rooted in an observation that *not all identity narratives are good for you*. Narrative scholars have been long comparing well-being-promoting and autonomy-promoting and -undermining narratives (Brody 1994; Farkas 2013;

Frank 1995; Morris 2012; Sontag 1978), both in medicine and psychiatry. Even if we were to accept the strong ethical narrative thesis that narrative is necessary for a flourishing life, recognizing the deleterious aspects of some narratives opens the door to the conclusion that one might be better off in a state of narrative-less existential puzzlement rather than stuck in the rut of a harmful narrative.

Implications of the Narrative Approach to Identity for the Assessment of DBS

Claims about what identity is all about are an interesting area for both personal reflection and philosophical inquiry, but we should perhaps maintain an amount of normative liberalism when transferring those debates into the applied sphere of healthcare provision. Although Schechtman's claims seem to be too strong to ground a strong blanket objection to the use of DBS, the narrative approach might help us to explore how the issues of identity mediate the translation of side effects and symptomatic improvement into changes in well-being.

Both illness and treatment may cause substantial change and disruptions in peoples' lives. As patients come to terms with their illness, creating illness narratives (Frank 1995; Phillips 2003) could enhance the sense of self and agency if it enables the agent to make sense of their experiences, including the experience of illness (Mackenzie and Poltera 2010). Treatment can disrupt previously attained equilibrium and may profoundly affect patients, their spouses, and their broader social environment (Agid et al. 2006; Schupbach et al. 2006).

Since reports of patients' understandings of the meaning of biomedically induced personality and behavioral changes perhaps most famously in reported in Kramer's *Listening to Prozac* (1994), it is clear that the influence of an illness, treatment, and side effects on patients' well-being is shaped by patients' ways of constructing an understanding of who they are, their openness to change of personality, ability, and relations to others, as well as perspectives on biomedical means of treatment. Those issues have not received broad attention despite the fact that, as DBS treatment is extended to neuropsychiatric disorders, we might anticipate their impact on the psychosocial adjustment of patients (Bell et al. 2009).

Conclusion and Further Directions

The application of DBS for neuropsychiatric disorders and enhancement is controversial, despite promising results of clinical trials for several conditions and the prospect of helping patients for whom nothing else has worked. The rhetorical evocations of the history of psychosurgery, the concepts of invasiveness and reversibility, and the therapy/enhancement distinction may unjustifiably bias the examination of costs, benefits, and moral assessment of the purposes that DBS could be legitimately put to. On the other hand, objections to the use of DBS are a good starting point for clarifying how the technology can serve us best. Worries

about mind control, although weak as an objection to the current use of DBS, draw our attention to the importance of considering dual-use, regulating with one eye on the future of DBS and including the consideration of users' liberty to choose and modify stimulation. Objections from personal and narrative identity are not strong enough to warrant halting the use of DBS, but further research could use narrative approaches to understand and improve the way DBS impacts the life of users.

Cross-References

▶ Biosecurity as a Normative Challenge
▶ Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
▶ Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine
▶ Dissociative Identity Disorder and Narrative
▶ Ethical Implications of Brain Stimulation
▶ Ethics in Neurosurgery
▶ Ethics of Functional Neurosurgery
▶ Ethics of Pharmacological Mood Enhancement
▶ Impact of Brain Interventions on Personal Identity
▶ Informed Consent and the History of Modern Neurosurgery
▶ Mind Reading, Lie Detection, and Privacy
▶ Neuroenhancement
▶ Neuroethics and Identity
▶ Neurosurgery: Past, Present, and Future
▶ Nonrestraint, Shock Therapies, and Brain Stimulation Approaches: Patient Autonomy and the Emergence of Modern Neuropsychiatry
▶ Parkinson's Disease and Movement Disorders – Historical and Ethical Perspectives
▶ Reflections on Neuroenhancement
▶ Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia
▶ The Morality of Moral Neuroenhancement
▶ Weaponization of Neuroscience

References

Agid, Y., Schupbach, M., Gargiulo, M., Mallet, L., Houeto, J., Behar, C., & Welter, M. L. (2006). Neurosurgery in Parkinson's disease: The doctor is happy, the patient less so? *Journal of Neural Transmission, 70*, s409–s414.

Andrade, P., Nobless, L. H. M., Temel, Y., Ackermans, L., Lim, L. W., Steinbusch, H. W. M., & Visser-Vandewalle, V. (2010). Neurostimulatory and ablative treatment options in major depressive disorder: A systematic review. *Acta Neurochirurgica, 152*, 565–577.

Baker, L. R. (2000). *Persons and bodies: A constitution view*. Cambridge, UK: Cambridge University Press.

Bauer, R., Pohl, S., Klosterkotter, J., & Kuhn, J. (2008). Deep brain stimulation in the context of addiction – A literature-based systematic evaluation. *Fortschritte der Neurologie – Psychiatrie, 76*, 396–401.

Baumeister, A. A. (2000). The Tulane electrical brain stimulation program – A historical case study in medical ethics. *Journal of the History of the Neurosciences, 9*(3), 262–278.

Baylis, F. (2011). "I am who I am": On the perceived threats to personal identity from deep brain stimulation. *Neuroethics*. doi:10.1007/s12152-011-9137-1.

Bell, E., & Racine, E. (2012). Ethical guidance for the use of deep drain stimulation in psychiatric trials and emerging uses: Review and reflections. In D. Denys, M. Feenstra, & R. Shuurman (Eds.), *Deep brain stimulation: A new frontier in psychiatry* (pp. 273–288). Berlin, Germany: Springer.

Bell, E., Mathieu, G., & Racine, E. (2009). Preparing the ethical future of deep brain stimulation. *Surgical Neurology, 72*(6), 577–586.

Brody, H. (1994). "My story is broken; can you help me fix it?": Medical ethics and the joint construction of narrative. *Literature and Medicine, 13*(1), 79–92.

Csoka, A. B., & Shipko, S. (2006). Persistent sexual side effects after SSRI discontinuation. *Psychotherapy and Psychosomatics, 75*(3), 187–188.

DeGrazia, D. (2005a). Enhancement technologies and human identity. *Journal of Medicine and Philosophy, 30*(3), 261–283.

DeGrazia, D. (2005b). *Human identity and bioethics*. Cambridge, UK: Cambridge University Press.

Edwards, T. C., Zrinzo, L., Limousin, P., & Foltynie, T. (2012). Deep brain stimulation in the treatment of chorea. *Movement Disorders, 27*, 357–363.

Farkas, C.-A. (2013). Potentially harmful side-effects: Medically unexplained symptoms, somatization, and the insufficient illness narrative for viewers of mystery diagnosis. *Journal of Medical Humanities, 34*(3), 315–328. doi:10.1007/s10912-013-9234-8.

Focquaert, F., & DeRidder, D. (2009). Direct intervention in the brain: Ethical issues concerning personal identity. *Journal of Ethics in Mental Health, 4*(2), 1–7.

Frank, A. (1995). *The wounded storyteller: Body, illness, and ethics*. Chicago, IL: University of Chicago Press.

Glannon, W. (2009). Stimulating brains, altering minds. *Journal of Medical Ethics, 35*, 289–292.

Greenberg, B. D., Nuttin, B., & Rezai, A. R. (2006). Education and neuromodulation for psychiatric disorders: A perspective for practitioners. *Neurosurgery, 59*(4), 717–719.

Halpern, C. H., Wolf, J. A., Bale, T. L., Stunkard, A. J., Danish, S. F., Grossman, M., & Baltuch, G. H. (2008). Deep brain stimulation in the treatment of obesity: A review. *Journal of Neurosurgery, 109*(10), 625–634.

Hariz, M. I., & Hariz, G. M. (2012). Hyping deep brain stimulation in psychiatry could lead to its demise. *British Medical Journal, 345*, e5447.

Hariz, M. I., Blomstedt, P., & Zrinzo, L. (2010). Deep brain stimulation between 1947 and 1987: The untold story. *Neurosurgical Focus, 29*(2), E1.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*(2), 102–111.

Hermans, H. J. M. (2003). The construction and reconstruction of a dialogical self. *Journal of Constructivist Psychology, 16*, 89–130.

Hermans, H. J. M., & Dimaggio, G. (Eds.). (2004). *The dialogical self in psychotherapy*. New York: Brunner-Routledge.

Hernando, V., Pastor, J., Pedrosa, M., Pena, E., & Sola, R. G. (2008). Low-frequency bilateral hypothalamic stimulation for treatment of drug-resistant aggressiveness in a young man with mental retardation. *Stereotactic and Functional Neurosurgery, 86*(4), 219–223.

Hildt, E. (2006). Electrodes in the brain: Some anthropological and ethical aspects of deep brain stimulation. *International Review of Information Ethics, 5*, 33–39.

Holtzheimer, P. E., Kelley, M. E., Gross, R. E., Filkowski, M. M., Garlow, S. J., Barrocas, A., & Mayberg, H. S. (2012). Subcallosal cingulate deep brain stimulation for treatment-resistent unipolar and bipolar depression. *Archives of General Psychiatry, 69*(2), 150–158.

Invasive. (n.d.). In Merriam-Webster online. Retrieved from http://www.merriam-webster.com/dictionary/invasive. Accessed 26 Nov 2012.

Krack, P., Hariz, M. I., Baunez, C., Guridi, J., & Obeso, J. A. (2010). Deep brain stimulation: From neurology to psychiatry? *Trends in Neurosciences, 33*, 474–484.

Kramer, P. D. (1994). *Listening to prozac*. London, UK: Fourth Estate Paperbacks.

Kuhn, J., Lenartz, D., Huff, W., Lee, S., Koulousakis, A., Klosterkoetter, J., & Sturm, V. (2007). Remission of alcohol dependency following deep brain stimulation of the nucleus accumbens: valuable therapeutic implications? *Journal of Neurology, Neurosurgery, and Psychiatry, 78*, 1152–1153.

Kuhn, J., Lenartz, D., Mai, J. K., Huff, W., Klosterkoetter, J., & Sturm, V. (2008). Disappearance of self-aggressive behavior in a brain-injured patient after deep brain stimulation of the hypothalamus: Technical case report. *Neurosurgery, 62*(5), e1182.

Kuhn, J., Gaebel, W., Klosterkoetter, J., & Woopen, C. (2009). Deep brain stimulation as a new therapeutic approach in therapy-resistant mental disorders: Ethical aspects of investigational treatment. *European Archives of Psychiatry and Clinical Neuroscience, 259*(2), s135–s141.

Kuhn, J., Gründler, T. O. J., Lenartz, D., Sturm, V., Klosterkötter, J., & Huff, W. (2010). Deep brain stimulation for psychiatric disorders (trans: Roseveare D.). *Deutsches Ärzteblatt International, 107*(7), 105–113.

Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Zumsteg, D., Wennberg, R., Keren, R., & Lozano, A. M. (2010). A phase 1 trial of deep brain stimulation of memory circuits in Alzheimer's disease. *Annals of Neurology, 68*(4), 521–534.

Lipsman, N., Zenar, R., & Bernstein, M. (2009). Personal identity, enhancement and neurosurgery: A qualitative study in applied neuroethics. *Bioethics, 23*(6), 375–383.

Lipsman, N., McAndrews, M. P., Lozano, A., & Bernstein, M. (2011). Research consent for deep brain stimulation in treatment-resistant depression: Balancing risk with patient expectations. *AJOB Neuroscience, 2*(1), 39–41.

Lipsman, N., Woodside, D. B., Giacobbe, P., Hamani, C., Carter, C., Norwood, S.J., Sutandar, K., Staab, R., Elias, G., Lyman, C. H., Smith, G. S., & Lozano, A. M. (2013). Subcallosal cingulate deep brain stimulation for treatment-refractory anorexia nervosa: a phase 1 pilot trial. *The Lancet, 381*(9875), 1361–1370.

Locke, J. (1975 [1694]). Book II, chapter XXVII, of identity and diversity. In P. H. Nidditch, (Ed.). *An essay concerning human understanding*. Oxford, UK: Clarendon Press.

Lozano, A. M. (2012). Deep brain stimulation therapy. *British Medical Journal, 344*, e1100.

Lu, L., Wang, X., & Kosten, T. R. (2009). Stereotactic neurosurgical treatment of drug addiction. *American Journal of Drug and Alcohol Abuse, 35*(6), 391–393.

Maan, A. K. (2010). *Internarrative identity: Placing the self* (2nd ed.). Lanham, MD: University Press of America.

Mackenzie, C., & Poltera, J. (2010). Narrative integration, fragmented selves, and autonomy. *Hypatia, 25*(1), 31–54.

Mathews, D. J. H., Rabins, P. V., & Greenberg, B. D. (2011). Deep brain stimulation for treatment-resistant neuropsychiatric disorders. In J. Illes & B. J. Sahakian (Eds.), *Oxford handbook of neuroethics* (pp. 441–453). Oxford, UK: Oxford University Press.

McMahan, J. (2002). *The ethics of killing: Problems at the margins of life*. Oxford, UK: Oxford University Press.

Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B., & Rosahl, S. (2007). *Intervening in the brain: Changing psyche and society*. Berlin, Germany: Springer.

Morris, D. B. (2012). Narrative and pain: Towards an integrative model. In R. J. Moore (Ed.), *Handbook of pain and palliative care* (pp. 733–751). New York: Springer.

Müller, S., & Christen, M. (2011). Deep brain stimulation in parkinsonian patients—ethical evaluation of cognitive, affective, and behavioral sequelae, AJOB Neuroscience, 2(1), 3–13.

Muzak, J. (2007). “They say the disease is responsible”: Social identity and the disease concept of drug addiction. In V. Raoul, C. Canam, A. D. Henderson, & C. Paterson (Eds.), *Unfitting stories: Narrative approaches to disease, disability, and trauma* (pp. 255–264). Waterloo, Canada: Wilfred Laurier Press.

Olson, E. T. (2003). Personal identity. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 352–368). Malden, MA: Blackwell.

Oshima, H., & Katayama, Y. (2010). Neuroethics of deep brain stimulation for mental disorders: Brain stimulation reward in humans. *Neurologia Medico-Chirurgica, 50*, 845–852.

Pacholczyk, A. (2011). DBS makes you feel good! – Why some of the ethical objections to the use of DBS for neuropsychiatric disorders and enhancement are not convincing. *Frontiers in Integrative Neuroscience, 5*(14).

Parfit, D. (1984). *Reasons and persons*. Oxford, UK: Clarendon.

Perry, J. (1972). Can the self divide? *The Journal of Philosophy, 69*(16), 463–488.

Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy, 25*(3), 162–177.

Phillips, J. (2003). Psychopathology and the narrative self. *Philosophy, Psychiatry and Psychology, 10*(4), 313–328.

Pressman, J. D. (1998). *Last resort: Psychosurgery and the limits of medicine*. Cambridge, UK: Cambridge University Press.

Racine, E., Waldman, S., Palmour, N., Risse, D., & Illes, J. (2007). “Currents of Hope”: Neurostimulation techniques in U.S. and U.K. print media. *Cambridge Quarterly of Healthcare Ethics, 16*, 312–316.

Ricoeur, P. (1984). *Time and narrative* (Vol. 1). Chicago: University of Chicago Press.

Ricoeur, P. (1985). *Time and narrative* (Vol. 2). Chicago: University of Chicago Press.

Ricoeur, P. (1988). *Time and narrative* (Vol. 3). Chicago: University of Chicago Press.

Ricoeur, P. (1992). *Oneself as another*. Chicago: University of Chicago Press.

Schechtman, M. (1996). *The constitution of selves*. Ithaca, NY: Cornell University Press.

Schechtman, M. (2009). Getting our stories straight: Self-narrative and personal identity. In D. J. H. Mathews, H. Bok, & P. V. Rabins (Eds.), *Personal identity and fractured selves* (pp. 65–92). Baltimore: Johns Hopkins University Press.

Schechtman, M. (2010). Philosophical reflections on narrative and deep brain stimulation. *Journal of Clinical Ethics, 21*(2), 133–139.

Schermer, M. (2011). Ethical issues in deep brain stimulation. *Frontiers in Integrative Neuroscience, 5*(17).

Schiff, N. D., Giacino, J. T., Kalmar, K., Victor, J. D., Baker, K., Gerber, M., & Rezai, A. R. (2007). Behavioural improvements with thalamic stimulation after severe traumatic brain injury. *Nature, 448*, 600–603.

Schuepbach, W. M. M., Rau, J., Knudsen, K., Volkmann, P., Krack, L., Timmermann, T. D., . . . EARLYSTIM Study Group. (2013). Neurostimulation for Parkinson’s disease with early motor complications. New England Journal of Medicine, *368*(7), 610–622.

Schupbach, M., Gargiulo, M., Welter, M. L., Mallet, L., Behar, C., Houeto, J. L., & Agid, Y. (2006). Neurosurgery in Parkinson disease: A distressed mind in a repaired body? *Neurology, 66*(12), 1811–1816.

Sententia, W. (2004). Cognitive liberty and converging technologies for improving human cognition. *Annals of the New York Academy of Science, 1013*, 221–228.

Smith, G. S., Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Diaconescu, A. O., Workman, C. I., & Lozano, A. M. (2012). Increased cerebral metabolism after 1 year of deep brain stimulation in Alzheimer disease. *Archives of Neurology, 69*(9), 1141–1148.

Sontag, S. (1978). *Illness as metaphor*. New York: Farrar, Straus and Giroux.

Strawson, G. (2004). Against narrativity. *Ratio, 17*(4), 428–452.

Synofzik, M., & Schlaepfer, T. E. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*(12), 1511–1520.

Synofzik, M., & Schlaepfer, T. E. (2011). Electrodes in the brain – Ethical criteria for research and treatment with deep brain stimulation for neuropsychiatric disorders. *Brain Stimulation, 4*, 7–16.

Synofzik, M., Schlaepfer, T. E., & Fins, J. J. (2012). How happy is too happy? Euphoria, neuroethics, and deep brain stimulation of the nucleus accumbens. *American Journal of Bioethics Neuroscience, 3*(1), 30–36.

Unger, P. (1990). *Identity, consciousness, and value*. Oxford, UK: Oxford University Press.

Vice, S. (2003). Literature and the narrative self. *Philosophy, 78*(1), 93–108.

Witt, K., Kuhn, J., Timmermann, L., Zurowski, M., & Woopen, C. (2011). Deep brain stimulation and the search for identity. *Neuroethics*. doi:10.1007/s12152-011-9100-1.

Yamamoto, T., Katayama, Y., Kobayashi, K., Oshima, H., Fukaya, C., & Tsubakawa, T. (2010). Deep brain stimulation for treatment of vegetative state. *European Journal of Neuroscience, 32*(7), 1145–1151.

Further Reading

Baerger, D. R., & McAdams, D. P. (1999). Life story coherence and its relation to psychological well-being. *Narrative Inquiry, 9*, 69–96.

Baumeister, R. F., & Newman, L. S. (1994). How stories make sense of personal experiences: Motives that shape autobiographical narratives. *Personality and Social Psychology Bulletin, 20*, 676–690.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology, 79*, 941–952. – and what about conversation partners during the decision making process and change.

Davies, B., & Harré, R. (1990). Positioning: The discursive production of selves. *Journal for the Theory of Social Behaviour, 20*, 43–63.

Hawkins, A. H. (1999). *Reconstructing illness: Studies in pathography*. West Lafayette, IN: Purdue University Press.

Hermans, H. J. M., & Hermans-Jansen, E. (1995). *Self-narratives: The construction of meaning in psychotherapy*. New York: Guilford Press.

Holstein, J. A., & Gubrium, J. F. (2000). *The self we live by: Narrative identity in a postmodern world*. New York: Oxford University Press.

Hydén, L. (1997). Illness and narrative. *Sociology of Health and Illness, 19*, 48–69.

Kleinman, A. (1988). *The illness narrative. Suffering, healing and the human condition*. New York: Basic Books.

Linde, C. (1993). *Life stories*. Oxford, UK: Oxford University Press.

Lodi-Smith, J., Geise, A. C., Roberts, B. W., & Robins, R. W. (2009). Narrating personality change. *Journal of Personality and Social Psychology, 96*(3), 679–689.

McAdams, D. (2001). The psychology of life stories. *Review of General Psychology, 5*(2), 100–122.

Pals, J. L. (2006). Narrative identity processing of difficult life experiences: Pathways of personality development and positive self-transformation in adulthood. *Journal of Personality, 74*(4), 1079–1110.

Strawson, G. (2007). Episodic ethics. In D. Hutto (Ed.), *Narrative and understanding persons*. Cambridge, UK: Cambridge University Press.

Section VIII

Ethical Implications of Brain Imaging

Neuroimaging Neuroethics: Introduction

40

Adina L. Roskies

Contents

Abstract

This section addresses some prominent neuroethical issues raised by new neuroimaging technologies. These include the potential for misinterpretation of neuroimages; combating this requires a better understanding of the difficulties and pitfalls in interpretation. Applications of neuroimaging introduce other ethical considerations. Discussed are the ethical implications of the emerging ability to assess disorders of consciousness with imaging in clinical settings, privacy concerns raised by the potential for mindreading; lie detection; and the ethical conundrums that will face us with the emerging potential for neuroimaging to provide predictive information about people's future behavior and disease.

Introduction

The ability to image brain structure and function in the intact, living, and behaving human raises a number of neuroethical questions regarding the proper interpretation, use, and limits of these new imaging technologies. Some questions are the

A.L. Roskies
Department of Philosophy, Dartmouth College, Hanover, NH, USA
e-mail: adina.roskies@dartmouth.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_34,

same or very similar to questions raised by other technologies, some are old questions in new guises, and some are specific to the technological and scientific challenges of neuroimaging. This section aims to address a spectrum of these neuroethical issues, from issues regarding interpretation, to those raised by potential applications.

Overview

Functional neuroimaging with fMRI is a highly technical, data-rich, and statistically complex endeavor. Unlike most highly technical and expensive scientific research, the technological sophistication of neuroimaging is almost paradoxically paired with an unprecedented ease of access to virtually anyone in academia or medicine. Its wide availability and ease of generating data mask the difficulty inherent in designing insightful and interpretable experiments. Good and fruitful neuroimaging studies are not trivial to design, and even these are challenging to interpret. Ethical applications of neuroimaging are predicated on sound interpretation. Thus, it is critical that researchers in neuroimaging and those applying neuroimaging results to real-world problems must be sophisticated producers and consumers of neuroimaging data. To that end, any complete neuroethical discussion of the area ought to include a tutorial on methods and pitfalls of neuroimaging research.

The ethical implications of neuroimaging abound. Standard use of neuroimaging for research purposes sometimes reveals brain abnormalities or other incidental findings with potentially grave consequences for the subject. Researchers must be aware of the statistical basis for interpreting the accuracy of such findings, as well as know the best procedures for informing the subject about these findings, with both her physical and psychological well-being in mind.

Neuroimaging has ethical implications not only in research studies, but in clinical medicine as well. Neuroimaging is poised to revolutionize some areas of medicine. Disorders of consciousness affect hundreds of thousands of people, and until recently diagnoses were made purely on a behavioral basis, with some people characterized as in a minimally conscious state, and others in a persistent vegetative state. Misdiagnosis of these conditions is pervasive (by some estimates, 37–43 %), and can have profound implications for medical strategies, pain management, and end of life decisions (Schnakers et al. 2009). As Demertzi and Laureys relate in their chapter, ▶ Chap. 41, "Detecting Levels of Consciousness," recent neuroimaging studies have provided novel tools for diagnosis and management of patients with altered states of consciousness. Perhaps most exciting is the potential avenue that brain imaging technologies provide to communicate with patients who are aware and cognitively intact, yet unable to respond verbally or behaviorally to instructions. Neuroimaging opens up the possibility that communication can be restored with these patients through associating responses with different imaginative tasks that have distinct neural signatures. The potential power of this mind-reading technique is coupled with a set of unique ethical challenges regarding how to interpret the capacity for autonomous and potentially momentous decision-making in patients whose abilities and lucidity are impaired to uncertain degrees and fluctuate over time.

Functional neuroimaging gives researchers some ability to infer aspects of mental state or behavior from brain activation patterns. Since all thoughts are due to brain activity, in principle it should be possible to correlate different thought contents with details of brain activation. In practice this is not possible except at a relatively coarse-grained level. The reasons for this are several: the spatial and temporal resolution of neuroimaging data is too low to capture the full spatial and temporal detail of brain dynamics; the signal-to-noise ratio is low. Researchers usually average multiple trials together to combat the noise in the data, thus losing some detail. Individual variability in brain structure and function suggest that no two brains will have identical structure-function relationships, which would mean that activity-meaning correlations would have to be mapped individually. Despite these limitations, there are also regularities evident across many subjects that can be exploited in order to allow some inferences about mental content (Roskies forthcoming). In recent years, headway has been made in understanding representations of semantic meaning for nouns, creating generative models, and reconstructing visual content from brain activation by perceptual stimuli. The limitations of imaging techniques provide reason to be skeptical of the prospect of being able to read propositional thought contents of people's minds, and very skeptical about abilities to do so covertly or without consent of the subject. On the other hand, there are less specific inferences one could make about content that could still raise ethical questions about the importance and limits of privacy. Roskies discusses some of these privacy issues in the chapter on Mind-reading.

Neuroimaging provides some traction on the perennial question of whether a person is lying or telling the truth. William Marsden, the inventor of the polygraph, was also the creator of Wonder Woman, the comic book heroine who had a golden lasso that could elicit truth-telling form those caught in its coils. People frequently think that neuroimaging has the promise to be a golden lasso – that we will be able to distinguish deceptive from honest answers by looking at brain activation data. Although neuroimaging has identified a constellation of brain areas that are reliably more active during lying than truth-telling in laboratory tests, no brain region is active in all such tests, and none that alone are sufficient for distinguishing honest from deceptive answers. Moreover, the vast majority of experiments meant to demonstrate the ability of neuroimaging to identify lying have used paradigms of questionable ecological validity – that is, they are investigating behaviors that are importantly unlike real-world cases of lying. Finally, even though a few experiments claim to have high rates of success in these discriminations, the difficulty in establishing error rates and the absence of base-rate information hampers application of these studies. Without known error rates and an understanding of the prevalence of lying in the relevant population, we cannot determine the actual reliability of these tests. Thus, despite promising results in laboratory settings, use in real-world applications from screening to forensic contexts seems premature. In the chapter on Mind-reading and privacy, Roskies touches on the problems plaguing real-world application of lie-detection studies.

Neuroimaging studies correlate brain activation patterns with tasks, behaviors, or phenotypes. Assuming that the sample upon which the correlations are based is

a representative one, one could potentially use neuroimaging predictively or diagnostically. Indeed, recent studies show that neuroimaging holds promise for diagnosing certain diseases, and for predicting recidivism in a penal population. The ability to predict behaviors and disease states brings with it a host of knotty ethical problems. For example, in the wrong hands diagnostic information could be used to discriminate against potential hires, insurance consumers, etc. Perhaps more disturbingly, one could imagine pressure to act upon predictions of future dangerousness, even in cases in which the subject has not exhibited any violent behavior. The shadow of "Minority Report", a movie in which prevision licensed arrest of future perpetrators before they committed crimes, looms large. It goes against our deep commitment to the principle that people are innocent until proven guilty, and pits our science against instrumental reasons for averting future harms. However, it is essential to realize that these methods only give statistical reasons for prediction, and that in any given case the predictions could be misleading. Not everyone with a genetic predisposition for a trait exhibits that trait, and because of statistical issues and complexities of brain organization and function, the same can be said of brain data. A discussion of the neuroethics of neuroimaging for prediction must address these complexities.

Conclusion and Future Directions

Neuroimaging is a fast-moving and evolving field, driven as much by advances in data analysis as by biological insight. No doubt as imaging techniques become more powerful and our picture of brain organization and function becomes better articulated, ethical issues will arise as practical realities that now seem only hypothetical, and perhaps entirely novel issues will raise their heads. For now, neuroimaging poses significant enough challenges, challenges that need to be addressed by technically-informed, philosophically sophisticated, and clear-minded analysis.

Cross-References

► A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience
► A Duty to Remember, a Right to Forget? Memory Manipulations and the Law
► Brain Research on Morality and Cognition
► Cognitive Liberty or the International Human Right to Freedom of Thought
► Consciousness and Agency
► Detecting Levels of Consciousness
► Drug Addiction and Criminal Responsibility
► Free Will, Agency, and the Cultural, Reflexive Brain
► Human Brain Research and Ethics
► Impact of Brain Interventions on Personal Identity
► Informed Consent and the History of Modern Neurosurgery

▶ Justice: A Neuroanthropological Account
▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
▶ Mind Reading, Lie Detection, and Privacy
▶ Neuroenhancement
▶ Neuroimaging and Criminal Law
▶ Neurolaw: Introduction
▶ Neuroscience, Neuroethics, and the Media
▶ Prediction of Antisocial Behavior
▶ The Morality of Moral Neuroenhancement
▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering
▶ The Use of Brain Interventions in Offender Rehabilitation Programs: Should It Be Mandatory, Voluntary, or Prohibited?

References

Roskies, A. L. (In press). Mind reading and privacy. In M. Gazzaniga (Ed.), *The cognitive neurosciences* V. Cambridge MA: MIT Press.

Schnakers, C., Faymonville, M.-E., & Laureys, S. (2009). Ethical implications: Pain, coma, and related disorders. In W. P. Banks (Ed.), *Encyclopedia of consciousness* (pp. 243–250). Oxford: Elsevier.

Detecting Levels of Consciousness

41

Athena Demertzi and Steven Laureys

Contents

Abstract

Patients with disordered consciousness due to brain injury pose medical and ethical challenges. Rates of clinical misdiagnosis of "vegetative"/unresponsive, minimally conscious and locked-in syndrome states remain challengingly high. Clinical misdiagnosis raises profound ethical concerns in terms of medical management, treatment of pain, and end-of-life decisions. Therefore, valid diagnosis is of utmost importance in clinical settings. A number of neuroimaging and electrophysiology studies now suggest that some behaviorally "vegetative state" patients may nevertheless show atypical cortical activation during resting state conditions; in some cases, they are able to follow commands or even communicate through willfully modified brain activity. Advances in investigating disorders of consciousness with neuroimaging techniques promise to lead to a more accurate understanding of individual patients' cognitive abilities and to shed light on the gray zones of these clinical conditions.

A. Demertzi (✉) • S. Laureys
Coma Science Group, Cyclotron Research Center & Neurology Department,
University of Liège, Liège, Belgium
e-mail: a.demertzi@ulg.ac.be

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_39,

The formulation of an ethical framework which will strike a balance between the protection of these patients and further research on disorders of consciousness is an ethical, clinical, and scientific demand.

Introduction

Consciousness is a multifaceted term for which there is no universal definition (Zeman 2001). Clinical experience teaches that we can define consciousness operationally, by reducing it to two components: wakefulness and awareness (Posner et al. 2007). Wakefulness refers to the level of alertness, and it is supported by the function of the subcortical arousal systems in the brainstem, the midbrain, and the thalamus. Clinically, wakefulness is indicated by eyes opening. Awareness refers to the content of consciousness, and is thought to be supported by the functional integrity of the cerebral cortex and its subcortical connections. Awareness can be further subdivided into awareness of environment and of self (James 1890). Clinically, awareness of environment is assessed by evaluating command-following and by observing nonreflex motor behavior, such as eye-tracking and localized responses to pain. Awareness of self, a more ill-defined concept, can be assessed by the patients' response to self-referential stimuli, such as the patients' own face in the mirror (Vanhaudenhuyse et al. 2008b). An illustrative example of the relationship between the two components of consciousness is the transition from full wakefulness to deep sleep: the less awake we are, the less aware we are of ourselves and our surroundings.

Developing technologies have made a significant impact on the assessment and management of patients with profoundly impaired consciousness. Neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), as well as electroencephalography (EEG) and evoked potential studies, have offered the possibility of objectively approaching cognitive processes in patients who are otherwise incapable of intelligible or sustained behavioral expression. Indeed, patients in a vegetative state (VS; Jennett and Plum 1972), recently coined unresponsive wakefulness syndrome (UWS; Laureys et al. 2010), do not show any signs of awareness of themselves or their environment in spite of preserved wakefulness levels as evidenced by eye opening. Some patients, however, may show minimal signs of consciousness but remain unable to reliably communicate with their environment. For this patient population, the term "minimally conscious state" (MCS) is used (Giacino et al. 2002). MCS was recently subcategorized based on the patients' complexity of behaviors (Bruno et al. 2012). More particularly, MCS + describes high-level behavioral responses, such as command-following, intelligible verbalizations or nonfunctional communication (i.e., responding to incorrectly or inconsistently to questions), and MCS − describes low-level behavioral responses, namely, visual pursuit, localization of noxious stimulation, or contingent behaviors like appropriate smiling or crying to emotional stimuli (e.g., patient cries only in her mother's presence). Patients for whom nonbehavioral evidence of consciousness or communication are only measurable via ancillary testing such as fMRI, PET, EEG, or evoked potentials can

be considered to be in a functional locked-in syndrome (LIS) (Bruno et al. 2011b). This is different from the classically defined LIS, where patients are fully conscious but their only means to communicate is by moving their eyes or their eyelids following a certain code, such as look up for "yes" and look down for "no" (American Congress of Rehabilitation Medicine 1995).

The Ethical Imperative

Early since disorders of consciousness appeared in the clinical setting, clinicians, scholars, theologians, and ethicists began to wonder what it is like to be in a state of profoundly disturbed consciousness (e.g., Thompson 1969). Are these unresponsive patients in pain, and can they even suffer from it? How can their quality of life be assessed? And more importantly, is a life in such severely restricted conditions worth living? Is this a question these patients can even weigh in on? Controversies of these kind mainly stem from how different people regard indefinite survival in disorders of consciousness (Jennett 2002; Demertzi et al. 2011a).

Despite the general view that quality of life is diminished in disease as a result of limited capacities to functionally engage in everyday living, these attitudes are formulated from a third-person perspective. Consequently, only rough estimations about what it is like to be in such a situation can be made. For instance, an analysis of public media reports on Terri Schiavo (a patient in a VS/UWS; e.g., Quill 2005) revealed that in some cases, the patient was described as feeling discomfort, which was incompatible with her clinical state (Racine et al. 2008). This implies that nonmedical individuals, whose opinions are supposedly represented – at least to a certain degree – by media reports, may be biased toward residual cognitive function of patients with consciousness alterations. Such bias could be attributed to the fact that patients' quality of life evaluations are made from the perspective of healthy individuals who tend to underestimate patients' subjective well-being (Demertzi et al. 2013b). Indeed, we recently showed that patients in LIS expressed a positive subjective quality of life contrary to what could be expected in this condition (Bruno et al. 2011a). As mentioned above, patients with LIS do not suffer from disorders of consciousness. However, LIS patients constitute a nice control population for patients with disorders of consciousness due to their resemblance in terms of physical disability and possibly common history (i.e., LIS patients can have been in comatose-like states). In this self-reporting survey, it was shown that the majority of patients in a chronic LIS, despite mentioning severe restrictions in community reintegration, professed good subjective well-being (Bruno et al. 2011a). The self-reported happiness status was associated with longer duration in this condition, namely, patients reported being happier over time, the ability to produce speech via assisting technologies, and lower rates of anxiety. These findings suggest that healthy persons who are not in direct contact with this patient population can have distorted pictures as to what is life in these severely constrained situations. But what about health-care workers opinions' who are more likely to interact with patients?

When clinicians were recently asked to express their opinions on possible pain perception in VS/UWS, a significant number of medical doctors ascribed pain perception in these patients (56 %) despite formal guidelines suggesting the opposite (e.g., The Multi-Society Task Force on PVS 1994). For MCS, there was no discrepancy in opinions and the majority (97 %) of respondents thought that MCS patients feel pain (Demertzi et al. 2009), in line with neuroimaging data strongly suggesting preserved pain perception in MCS (Boly et al. 2008). The issue of pain management in unresponsive patients becomes more challenging when withdrawal from life-supporting treatments, such as artificial nutrition and hydration, has been agreed upon. In a wide survey around Europe, we showed that health-care workers' opinions on end-of-life in disorders of consciousness differed depending on the diagnosis (i.e., respondents supported treatment withdrawal more often for VS/UWS than MCS patients), professional background (i.e., when physicians imagined being in MCS, they preferred more often to be kept alive compared to paramedical professionals), region of origin (i.e., Northern Europeans agreed with treatment withdrawal more often compared to Central and Southern European respondents), and religious beliefs (i.e., religious respondents agreed less with treatment limitation in both VS/UWS and MCS compared to nonreligious ones) (Demertzi et al. 2011b). These data show that personal opinions about ethical issues in disorders of consciousness differ and hence different clinical practice can be expected. For example, in the European survey, the majority of participants approved to stop treatment in VS/UWS (66 %) more than MCS (28 %). In this case, VS/UWS patients may run the risk of being left without administration of opioids or other analgesic drugs during their dying process (Laureys 2005; Fins 2006) on the grounds that they are unable to experience suffering due to hunger or thirst (Ahronheim and Gasner 1990). We recently showed an interaction between opinions on pain perception and end-of-life preferences. More particularly, there was a negative association between these two types of attitudes for the VS/UWS, namely, that participants agreed with treatment withdrawal when they denied these patients could experience pain (Demertzi et al. 2013a). From such studies on clinicians' attitudes and attitudes of patients' families (e.g., Kuehlmeyer et al. 2012), it becomes evident that medical and ethical controversies continue to exist for patients with disorders of consciousness. In order to resolve them, at least to a certain degree, we need to improve our current understanding of how these patients function. The use of objective biomarkers may help us to better determine the differences in underlying pathophysiology characterizing the clinical entities of consciousness. Consequently, clinicians are expected to learn about patients' values and preferences and to maintain clinical insight for changes in patient's status with the patients' best interests in mind (Jox et al. 2012).

The Pursuit of Awareness

The past 15 years have provided an unprecedented collection of discoveries that bear upon our scientific understanding of consciousness in the human brain

following severe brain damage. Highlighted among these discoveries are unique demonstrations that patients with little or no behavioral evidence of conscious awareness may retain critical cognitive capacities. These first scientific demonstrations support that some severely brain-injured patients in longstanding conditions of limited behavioral responsiveness may nonetheless retain latent capacities for awareness. Such capacities include the human functions of language and higher-level cognition that, either spontaneously or through direct interventions, may reemerge even at long-time intervals or can remain unrecognized.

The experimental protocols adopted by neuroimaging and electrophysiology studies to assess residual brain function in patients with disorders of consciousnesses include data acquisitions both in resting state conditions and after external stimulation. With appropriate experimental design, these technologies also permit detection of nonverbal command-following and may even establish methods of communication with some behaviorally unresponsive patients.

Resting State Passive Paradigms

In resting conditions, when we do not perform any task and receive no external stimulation, our brains are not inactive. It rather seems that in the resting state, there is some kind of typical cognitive activity when the mind is unconstrained and wanders (e.g., Mason et al. 2007; Raichle and Snyder 2007). We have recently characterized the phenomenological complexity of such cognitive process by two components: *external* awareness, namely, everything we perceive through our senses (what we see, hear, feel, smell and taste), and *internal* awareness or stimulus-independent thoughts (Demertzi et al. 2013c). In a combined behavioral-neuroimaging experimental setup (Vanhaudenhuyse et al. 2011), we first showed that external and internal awareness behavioral ratings were anticorrelated and that there was a switch in their dominance at a similar frequency as observed in the fMRI BOLD signal (Fransson 2005). Interestingly, the switch between the external and internal milieu was found not only to characterize overt behavioral reports but also had a cerebral correlate, More particularly, it was shown that behavioral reports of internal awareness were linked to the activity of midline anterior cingulate/mesiofrontal areas as well as posterior cingulate/precuneal cortices. Inversely, subjective ratings for external awareness correlated with the activity of lateral fronto-parieto-temporal regions. Such fMRI resting state studies show that a network of brain areas seems to play an important role in sustaining consciousness. Functional connectivity in the posterior cingulate cortex, medial prefrontal cortex, and posterior parietal cortices (these areas broadly known as the default mode network-DMN) show no functional connectivity in brain death (Boly et al. 2009; Soddu et al. 2011). Additionally, the functional connectivity pattern of this network was found to decrease as a function of the level of consciousness, ranging from controls and patients with locked-in syndrome to MCS, VS/UWS, and coma (Vanhaudenhuyse et al. 2010). More recently it was shown that more networks during resting state (Heine et al. 2012) show disrupted connectivity in patients,

which was able to discriminate them from healthy subjects with 85& accuracy (Demertzi et al. in press). Although the value of the fMRI resting state as a differential diagnostic tool remains to be determined, the potentially prognostic value of DMN connectivity was shown in a cohort of patients in the acute stage of coma for whom the presence of DMN functional connectivity was correlated with reversibility of coma. That is, preserved functional connectivity of the DMN was observed in those patients who subsequently regained consciousness and was disrupted in all patients who failed to recover their consciousness levels, suggesting that resting state DMN connectivity may serve as an indicator of the extent of cortical disruption (Norton et al. 2012).

PET studies in resting state show that unresponsive patients compared to healthy subjects are characterized by reduced levels of global metabolism. Interestingly, recovery from VS/UWS does not necessarily coincide with resumption of global metabolic activity (Laureys et al. 2004a). Rather, some areas appear to be more important than others for conscious function. For instance, patients in VS/UWS show impaired metabolism in a widespread network encompassing midline (i.e., anterior cingulate/mesiofrontal and posterior cingulate cortex/precuneus) and lateral (i.e., prefrontal and posterior parietal) associative cortices, compared to healthy controls (Laureys et al. 1999). Importantly, functional connections of these areas with the thalami were restored after recovery from VS/UWS (Laureys et al. 2000). Using PET, it was recently shown that VS/UWS patients compared to healthy controls exhibit metabolic dysfunction in both external and internal awareness networks as well as in the thalami. In contrast, MCS patients compared to healthy controls showed dysfunction mostly in internal awareness network and thalami, which could reflect an altered self-awareness in these patients that is difficult to quantify at the bedside (Thibaut et al. 2012).

Various EEG paradigms have also made an effort to differentiate between the clinical entities of disorders of consciousness. Fifteen-minute EEG resting state acquisitions showed that VS/UWS patients had significantly higher correlated oscillations than MCS in the delta frequency band (Lehembre et al. 2012). Such delta frequency activity is represented as a high amplitude brain wave with a oscillation between 0 Hz and 4 Hz usually associated with the deepest stages of sleep also known as slow-wave sleep (SWS). As previously shown, power in the delta band also increases with severity of disorders of consciousness (Leon-Carrion et al. 2008). Similarly, the bi-spectral index, a measure of the depth of anesthesia, was shown to discriminate between VS/UWS and MCS patients (Schnakers et al. 2005). The bi-spectral index was also positively correlated with behavioral scores of awareness at the time of testing and associated with outcome results at 1 year post-trauma. Additionally, an EEG entropy score of 52 (value ranging from 0 to 91, with higher scores indicating higher consciousness level) was shown to be able to differentiate acutely unconscious from MCS patients with 89 % sensitivity and 90 % specificity. However, the prognostic value of this measure was not high. At 1-year follow-up in the acute setting, patients with good outcome (i.e., recovery of functional communication) could be discriminated from those with bad outcomes (death or still in VS/UWS) with a specificity of 78 % and sensitivity of 60 % at the

entropy value of 57. Considering the moderate discrimination capacity of entropy measures for outcome this clinical feature cannot be recommended as a prognostic tool (Gosseries et al. 2011). Other efforts have also been made to use EEG signal patterns as a prognostic tool for these patients. For example, it has been observed that VS/UWS patients who made a behavioral recovery at the 3-month follow-up showed higher occipital source power in the alpha band of resting EEG when compared to those who did not (Babiloni et al. 2009). Normally, high power of pre-stimulus cortical alpha rhythms (about 8–12 Hz) underlies conscious perception in healthy subjects. As such, cortical sources of resting alpha rhythms might predict recovery in VS/UWS patients.

Taken together, resting state studies with fMRI, PET, and EEG provide promising techniques for assessing residual brain function in patients with disorders of consciousness, thanks to their simple and fast application and the fact that no collaboration is needed on patients' behalf. It should be noted, however, that in terms of differential diagnosis, no A-level recommendations can be made yet favoring a particular type of testing. Large multicentric studies are necessary to validate their diagnostic power and determine their prognostic value.

Activation Studies

Importantly, brain responses to external stimuli provide valuable information not only about the preserved functional (and to some degree anatomical) connectivity among distinct brain regions but also about the nature of detected responses. The potential for pain perception in MCS patients is suggested by findings of cerebral correlates of pain processing in a network similar to that in healthy controls (Boly et al. 2008). The activation pattern observed in MCS patients was also much more widespread than in VS/UWS patients, suggesting a difference in capacity for pain perception. The type of stimuli used also seems to make a difference to the observed neural responses and therefore further assists in the inference of awareness. Stimuli with emotional valence, such as infant cries and the patient's own name, induced a much more widespread activation in MCS patients than did meaningless noise (Laureys et al. 2004b). The activation pattern was comparable with that previously obtained in healthy controls. Patients also showed higher fMRI activity in the anterior cingulate cortex after listening to their own name as compared to listening to a familiar name, and this activity correlated with the behaviorally assessed level of consciousness of the patient (Qin et al. 2010). Such results imply that self-referential stimuli, like one's own name, are attention grabbing and therefore can be used in the assessment of residual brain function of these patients.

EEG studies measuring effective connectivity also seem to be able to differentiate between VS/UWS and MCS patients. Effective connectivity is a measure of the causal relationship between brain areas. One study using a mismatch negativity paradigm and applying dynamic causal modeling found that the only significant difference in functional connectivity between VS/UWS and MCS patients was an impairment of backward connectivity from frontal to temporal cortices (Boly et al. 2011).

Also, measurement of EEG effective connectivity after the application of transcranial magnetic stimulation (TMS) revealed that VS/UWS patients showed a simple, local response after the TMS pulses. In contrast, MCS patients showed more complex activations after the TMS pulses which involved distant cortical areas ipsilateral and contralateral to the site of stimulation (Rosanova et al. 2012). A number of studies have used event-related potentials (ERPs: averages of segments of EEG locked to the presentation of a stimulus) to assess patients with disordered consciousness. In healthy subjects, even though early "exogenous" components (elicited within 100 ms after stimulus presentation) are known to persist even in unconsciousness, later "endogenous" ERP components (e.g. P300, i.e., showing a positive peak 300 ms after stimulus presentation) can be used to infer conscious cognitive processing of information (Vanhaudenhuyse et al. 2008a). In several ERP studies, the detection of the patients' own name was used in order to assess residual linguistic preservation and self-processing. When their own name was presented infrequently among other names, a differential P300 has been observed in locked-in syndrome, MCS and VS/UWS patients, implying that the auditory system was relatively preserved in response to language stimulation (Schnakers et al. 2008, 2009a). This suggests high-level cognitive processing for auditory stimuli beyond their most basic features in patients with disorders of consciousness.

Active Paradigms

The primary challenge with many activation studies is that, in the absence of a subjective contribution from patients, a similar-to-control brain activation pattern cannot necessarily be interpreted as evidence of a conscious percept. In that respect, paradigms measuring command-following by means of mental imagery appear more promising for allowing inferences of awareness in behaviorally unresponsive patients. The first of such "active" mental imagery paradigms was developed using fMRI. Healthy volunteers were asked to perform a series of mental tasks (e.g., imagine singing a song in your head or imagine your mother's face). The most robust and distinguishable patterns of brain activation were obtained using motor mental imagery (i.e., imagine playing tennis) and spatial navigation (i.e., imaging walking around in your house). These two distinct tasks led to the predicted activation of supplementary motor cortex and parahippocampal areas, respectively (Boly et al. 2007). Initially, the use of this paradigm in a clinically diagnosed VS/UWS patient showed brain activation to both mental tasks indistinguishable from controls, suggesting preserved awareness in this patient preceding clinical evaluation (Owen et al. 2006). In a larger cohort of 54 patients clinically diagnosed as being either VS/UWS or MCS, five patients showed robust fMRI evidence of response to these two mental task commands (Monti et al. 2010). The best explanation for their ability to exhibit brain activity differentially responding to verbal commands is that they were consciously aware of, understood, and could comply with the instructions. In one patient, functional communication was established by explaining to the patient to do the motor imagery task to communicate "yes" and the spatial navigation task to

communicate "no." The automated user-independent analysis of the acquired fMRI data classified the brain's responses as a "yes" or "no" answer to a series of simple questions. Correct answers were obtained and reported by blinded examiners for five consecutive questions. Only for the final question could no determinate answer be elicited, due to absent brain activation. This patient was originally diagnosed as VS/UWS on the basis of standardized behavioral assessments, but the fMRI results showed that he was actually in an MCS (Monti et al. 2010). Thus, the fMRI experiment can identify instances of clinical misdiagnosis, which are probably frequent in clinical settings (Schnakers et al. 2009b). This patient could also be considered to be in a functional LIS, given that it was only functional neuroimaging that permitted the establishment of the yes-no communication to closed questions (Bruno et al. 2011b).

Unfortunately, such "active" paradigms currently can only be seen as a proof of concept rather than a practical communication tool. As soon as the patient was removed from the MRI machine, no further communication was possible. Hence, portable and cheaper EEG-based equivalents (e.g., Cruse et al. 2011, 2012; Lule et al. 2013) are being developed for more routine clinical use (for recent review see Chatelle et al. 2012). Such brain-computer interfaces (BCIs) have already been used successfully in clinical settings.

It is important to stress that the absence of brain activation to commands cannot be taken as proof of absence of consciousness. Repeated fMRI and EEG BCI assessments are needed to increase the confidence for true negative findings. There is also the problem of false-positives, namely, the fact that unconscious patients may show artifact or noise-related activation (Soddu et al. 2011). Future studies should deal with these issues in large patient cohorts and should also assess test-retest variability of these novel technologies in this specific context. What remains to be shown, however, is whether such technologies can be used as evidence of the expressed will of a competent patient (Gantner et al. 2012). For example, how can a negative response of an "unresponsive" patient to the question of whether they want to continue to live be considered as a reliable response to be respected? Similarly, should pain treatment in an MCS patient change once they communicate that they suffer? Should proving consciousness in these patients be considered as piece of evidence to be celebrated, or can it work against patients' and families' best interests? (Jox et al. 2012) These aforementioned questions require answers that future establishment of ethical and legal provisions can provide.

Conclusions and Future Directions

Using advanced neuroimaging and electrophysiology techniques, a number of clinically unresponsive patients may show remnants of preserved awareness, judging from their ability to follow commands and in some cases even communicate through their willfully modified brain activity. Studies of this kind highlight the pitfalls of forming a diagnosis based on behavioral assessments only. Using these techniques, the gray zones between the different disorders of consciousness in the

clinical spectrum following coma are beginning to be better defined (Laureys and Boly 2008). It should be stressed that these exciting developments are not yet routine in clinical practice. The first obstacle that needs to be overcome before the methodologies discussed above can enter clinical practice relates to ethical concerns. For example, in terms of treatment planning, such as pain management and end-of-life decision making, patients with disorders of consciousness are now offered the possibility to express their preferences by means of brain-computer interfaces. What remains to be clarified is the degree to which such indirect responses can be considered reliable and worthy of legal representation. It is expected that these methods will eventually help clinicians and other caregivers understand the state of disordered consciousness of their patients.

Cross-References

► Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges
► Ethical Implications of Brain–Computer Interfacing

References

Ahronheim, J. C., & Gasner, M. R. (1990). The sloganism of starvation. *Lancet, 335*, 278–279.

American Congress of Rehabilitation Medicine. (1995). Recommendations for use of uniform nomenclature pertinent to patients with severe alterations of consciousness. *Archives of Physical Medicine and Rehabilitation, 76*, 205–209.

Babiloni, C., Sara, M., Vecchio, F., Pistoia, F., Sebastiano, F., Onorati, P., Albertini, G., Pasqualetti, P., Cibelli, G., Buffo, P., & Rossini, P. M. (2009). Cortical sources of resting-state alpha rhythms are abnormal in persistent vegetative state patients. *Clinical Neurophysiology, 120*, 719–729.

Boly, M., Coleman, M. R., Davis, M. H., Hampshire, A., Bor, D., Moonen, G., Maquet, P. A., Pickard, J. D., Laureys, S., & Owen, A. M. (2007). When thoughts become action: An fMRI paradigm to study volitional brain activity in non-communicative brain injured patients. *NeuroImage, 36*, 979–992.

Boly, M., Faymonville, M.-E., Schnakers, C., Peigneux, P., Lambermont, B., Phillips, C., Lancellotti, P., Luxen, A., Lamy, M., Moonen, G., Maquet, P., & Laureys, S. (2008). Perception of pain in the minimally conscious state with PET activation: An observational study. *Lancet Neurology, 7*, 1013–1020.

Boly, M., Tshibanda, L., Vanhaudenhuyse, A., Noirhomme, Q., Schnakers, C., Ledoux, D., Boveroux, P., Garweg, C., Lambermont, B., Phillips, C., Luxen, A., Moonen, G., Bassetti, C., Maquet, P., & Laureys, S. (2009). Functional connectivity in the default network during resting state is preserved in a vegetative but not in a brain dead patient. *Human Brain Mapping, 30*, 2393–2400.

Boly, M., Garrido, M. I., Gosseries, O., Bruno, M. A., Boveroux, P., Schnakers, C., Massimini, M., Litvak, V., Laureys, S., & Friston, K. (2011). Preserved feedforward but impaired top-down processes in the vegetative state. *Science, 332*, 858–862.

Bruno, M.-A., Bernheim, J. L., Ledoux, D., Pellas, F., Demertzi, A., & Laureys, S. (2011a). A survey on self-assessed well-being in a cohort of chronic locked-in syndrome patients: Happy majority, miserable minority. *British Medical Journal Open, 1*, 1–9.

Bruno, M.-A., Vanhaudenhuyse, A., Thibaut, A., Moonen, G., & Laureys, S. (2011b). From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes:

Recent advances in our understanding of disorders of consciousness. *Journal of Neurology, 258*, 1373–1384.

Bruno, M. A., Majerus, S., Boly, M., Vanhaudenhuyse, A., Schnakers, C., Gosseries, O., Boveroux, P., Kirsch, M., Demertzi, A., Bernard, C., Hustinx, R., Moonen, G., & Laureys, S. (2012). Functional neuroanatomy underlying the clinical subcategorization of minimally conscious state patients. *Journal of Neurology, 259*, 1087–1098.

Chatelle, C., Chennu, S., Noirhomme, Q., Cruse, D., Owen, A. M., & Laureys, S. (2012). Brain-computer interfacing in disorders of consciousness. *Brain Injury, 26*, 1510–1522.

Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T. A., Fernandez-Espejo, D., Pickard, J. D., Laureys, S., & Owen, A. M. (2011). Bedside detection of awareness in the vegetative state: A cohort study. *Lancet, 378*, 2088–2094.

Cruse, D., Chennu, S., Chatelle, C., Fernandez-Espejo, D., Bekinschtein, T. A., Pickard, J. D., Laureys, S., & Owen, A. M. (2012). Relationship between etiology and covert cognition in the minimally conscious state. *Neurology, 78*, 816–822.

Demertzi, A., Schnakers, C., Ledoux, D., Chatelle, C., Bruno, M.-A., Vanhaudenhuyse, A., Boly, M., Moonen, G., & Laureys, S. (2009). Different beliefs about pain perception in the vegetative and minimally conscious states: A European survey of medical and paramedical professionals. *Progress in Brain Research, 177*, 329–338.

Demertzi, A., Laureys, S., & Bruno, M. A. (2011a). The ethics in disorders of consciousness. In J. L. Vincent (Ed.), *Annual update in intensive care and emergency medicine* (pp. 675–682). Berlin: Springer.

Demertzi, A., Ledoux, D., Bruno, M.-A., Vanhaudenhuyse, A., Gosseries, O., Soddu, A., Schnakers, C., Moonen, G., & Laureys, S. (2011b). Attitudes towards end-of-life issues in disorders of consciousness: A European survey. *Journal of Neurology, 258*, 1058–1065.

Demertzi, A., Racine, E., Bruno, M.-A., Ledoux, D., Gosseries, O., Vanhaudenhuyse, A., Thonnard, M., Soddu, A., Moonen, G., & Laureys, S. (2013a). Pain perception in disorders of consciousness: Neuroscience, clinical care, and ethics in dialogue. *Neuroethics, 6*(1), 37–50.

Demertzi, A., Gosseries, O., Ledoux, D., Laureys, S. & Bruno, M.-A. (2013b). Quality of life and end-of-life decisions after brain injury. In N. Warren & L. Manderson (Ed.), *Rethinking disability and quality of life: A global perspective* (pp. 95–110). Dordrecht: Springer.

Demertzi, A., Soddu, A., & Laureys, S. (2013c). Consciousness supporting networks. *Current Opinion in Neurobiology, 23*(2), 239–244.

Demertzi, A., Gómez, F., Crone, J. S., Vanhaudenhuyse, A., Tshibanda, L., Noirhomme, Q., Thonnard, M., Charland-Verville, V., Kirsch, M., Laureys, S., & Soddu, A. (in press). Multiple FMRI system-level baseline connectivity is disrupted in patients with consciousness alterations. Cortex.

Fins, J. J. (2006). Affirming the right to care, preserving the right to die: Disorders of consciousness and neuroethics after Schiavo. *Palliative and Supportive Care, 4*, 169–178.

Fransson, P. (2005). Spontaneous low-frequency BOLD signal fluctuations: An fMRI investigation of the resting-state default mode of brain function hypothesis. *Human Brain Mapping, 26*, 15–29.

Gantner, I. S., Bodart, O., Laureys, S., & Demertzi, A. (2012). Our rapidly changing understanding of acute and chronic disorders of consciousness: Challenges for neurologists. *Future Neurology, 8*, 43–54.

Giacino, J. T., Ashwal, S., Childs, N., Cranford, R., Jennett, B., Katz, D. I., Kelly, J. P., Rosenberg, J. H., Whyte, J., Zafonte, R. D., & Zasler, N. D. (2002). The minimally conscious state: Definition and diagnostic criteria. *Neurology, 58*, 349–353.

Gosseries, O., Schnakers, C., Ledoux, D., Vanhaudenhuyse, A., Bruno, M.-A., Demertzi, A., Noirhomme, Q., Lehembre, R., Damas, P., Goldman, S., Peeters, E., Moonen, G., & Laureys, S. (2011). Automated EEG entropy measurements in coma, vegetative state/unresponsive wakefulness syndrome and minimally conscious state. *Functional Neurology, 36*, 25–30.

Heine, L., Soddu, A., Gomez, F., Vanhaudenhuyse, A., Tshibanda, L., Thonnard, M., Charland-Verville, V., Kirsch, M., Laureys, S., & Demertzi, A. (2012). Resting state networks and

consciousness. Alterations of multiple resting state network connectivity in physiological, pharmacological and pathological consciousness states. *Frontiers in Psychology, 3*, 1–12.

James, W. (1890). Attention. In *The principles of psychology* (pp. 402–458). New York: Dover.

Jennett, B. (2002). Attitudes to the permanent vegetative state. In B. Jennett (Ed.), *The vegetative state. Medical facts, ethical and legal dilemmas* (pp. 97–125). Cambridge: Cambridge University Press.

Jennett, B., & Plum, F. (1972). Persistent vegetative state after brain damage. A syndrome in search of a name. *Lancet, 1*, 734–737.

Jox, R. J., Bernat, J. L., Laureys, S., & Racine, E. (2012). Disorders of consciousness: Responding to requests for novel diagnostic and therapeutic interventions. *Lancet Neurology, 11*, 732–738.

Kuehlmeyer, K., Borasio, G. D., & Jox, R. J. (2012). How family caregivers' medical and moral assumptions influence decision making for patients in the vegetative state: A qualitative interview study. *Journal of Medical Ethics, 38*, 332–337.

Laureys, S. (2005). Science and society: Death, unconsciousness and the brain. *Nature Review Neuroscience, 6*, 899–909.

Laureys, S., & Boly, M. (2008). The changing spectrum of coma. *Nature Clinical Practice Neurology, 4*, 544–546.

Laureys, S., Goldman, S., Phillips, C., Van Bogaert, P., Aerts, J., Luxen, A., Franck, G., & Maquet, P. (1999). Impaired effective cortical connectivity in vegetative state: Preliminary investigation using PET. *NeuroImage, 9*, 377–382.

Laureys, S., Faymonville, M. E., Luxen, A., Lamy, M., Franck, G., & Maquet, P. (2000). Restoration of thalamocortical connectivity after recovery from persistent vegetative state. *Lancet, 355*, 1790–1791.

Laureys, S., Owen, A. M., & Schiff, N. D. (2004a). Brain function in coma, vegetative state, and related disorders. *Lancet Neurology, 3*, 537–546.

Laureys, S., Perrin, F., Faymonville, M.-E., Schnakers, C., Boly, M., Bartsch, V., Majerus, S., Moonen, G., & Maquet, P. (2004b). Cerebral processing in the minimally conscious state. *Neurology, 63*, 916–918.

Laureys, S., Celesia, G. G., Cohadon, F., Lavrijsen, J., Leon-Carrrion, J., Sannita, W. G., Sazbon, L., Schmutzhard, E., von Wild, K. R., Zeman, A., Dolce, G., & European Task Force on Disorders of Consciousness. (2010). Unresponsive wakefulness syndrome: A new name for the vegetative state or apallic syndrome. *BMC Medicine, 8*, 68.

Lehembre, R., Marie-Aurelie, B., Vanhaudenhuyse, A., Chatelle, C., Cologan, V., Leclercq, Y., Soddu, A., Macq, B., Laureys, S., & Noirhomme, Q. (2012). Resting-state EEG study of comatose patients: A connectivity and frequency analysis to find differences between vegetative and minimally conscious states. *Functional Neurology, 27*, 41–47.

Leon-Carrion, J., Martin-Rodriguez, J. F., Damas-Lopez, J., Barroso y Martin, J. M., & Dominguez-Morales, M. R. (2008). Brain function in the minimally conscious state: A quantitative neurophysiological study. *Clinical Neurophysiology, 119*, 1506–1514.

Lule, D., Noirhomme, Q., Kleih, S. C., Chatelle, C., Halder, S., Demertzi, A., Bruno, M.-A., Gosseries, O., Vanhaudenhuyse, A., Schnakers, C., Thonnard, M., Soddu, A., Kubler, A., & Laureys, S. (2013). Probing command following in patients with disorders of consciousness using a brain-computer interface. *Clinical Neurophysiology, 124*(1), 101–116.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science, 315*, 393–395.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful modulation of brain activity in disorders of consciousness. *The New England Journal of Medicine, 362*, 579–589.

Norton, L., Hutchison, R. M., Young, G. B., Lee, D. H., Sharpe, M. D., & Mirsattari, S. M. (2012). Disruptions of functional connectivity in the default mode network of comatose patients. *Neurology, 78*, 175–181.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science, 313*, 1402.

Posner, J., Saper, C., Schiff, N. D., & Plum, F. (Eds.). (2007). *Plum and Posner's diagnosis of stupor and coma*. New York: Oxford University Press.

Qin, P., Di, H., Liu, Y., Yu, S., Gong, Q., Duncan, N., Weng, X., Laureys, S., & Northoff, G. (2010). Anterior cingulate activity and the self in disorders of consciousness. *Human Brain Mapping, 31*, 1993–2002.

Quill, T. E. (2005). Terri Schiavo–A tragedy compounded. *The New England Journal of Medicine, 352*, 1630–1633.

Racine, E., Amaram, R., Seidler, M., Karczewska, M., & Illes, J. (2008). Media coverage of the persistent vegetative state and end-of-life decision-making. *Neurology, 71*, 1027–1032.

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage, 37*, 1083–1090; discussion 1097–1089.

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M. A., Mariotti, M., Boveroux, P., Tononi, G., Laureys, S., & Massimini, M. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain, 135*, 1308–1320.

Schnakers, C., Majerus, S., & Laureys, S. (2005). Bispectral analysis of electroencephalogram signals during recovery from coma. *Neuropsychological Rehabilitation, 15*, 381–388.

Schnakers, C., Perrin, F., Schabus, M., Majerus, S., Ledoux, D., Damas, P., Boly, M., Vanhaudenhuyse, A., Bruno, M. A., Moonen, G., & Laureys, S. (2008). Voluntary brain processing in disorders of consciousness. *Neurology, 71*, 1614–1620.

Schnakers, C., Perrin, F., Schabus, M., Hustinx, R., Majerus, S., Moonen, G., Boly, M., Vanhaudenhuyse, A., Bruno, M. A. & Laureys, S. (2009a). Detecting consciousness in a total locked-in syndrome: An active event-related paradigm. *Neurocase, 15*, 271–277.

Schnakers, C., Vanhaudenhuyse, A., Giacino, J. T., Ventura, M., Boly, M., Majerus, S., Moonen, G., & Laureys, S. (2009b). Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neurobehavioral assessment. *BMC Neurology, 9*, 35.

Soddu, A., Vanhaudenhuyse, A., Demertzi, A., Marie-Aurelie, B., Tshibanda, L., Di, H., Melanie, B., Papa, M., Laureys, S., & Noirhomme, Q. (2011). Resting state activity in patients with disorders of consciousness. *Functional Neurology, 26*, 37–43.

The Multi-Society Task Force on PVS. (1994). Medical aspects of the persistent vegetative state (2). *The New England Journal of Medicine, 330*, 1572–1579.

Thibaut, A., Bruno, M. A., Chatelle, C., Gosseries, O., Vanhaudenhuyse, A., Demertzi, A., Schnakers, C., Thonnard, M., Charland-Verville, V., Bernard, C., Bahri, M., Phillips, C., Boly, M., Hustinx, R., & Laureys, S. (2012). Metabolic activity in external and internal awareness networks in severely brain-damaged patients. *Journal of Rehabilitation Medicine, 44*, 487–494.

Thompson, G. T. (1969). An appeal to doctors. *Lancet, 2*, 1353.

Vanhaudenhuyse, A., Laureys, S., & Perrin, F. (2008a). Cognitive event-related potentials in comatose and post-comatose states. *Neurocritical Care, 8*, 262–270.

Vanhaudenhuyse, A., Schnakers, C., Bredart, S., & Laureys, S. (2008b). Assessment of visual pursuit in post-comatose states: Use a mirror. *Journal of Neurology, Neurosurgery and Psychiatry, 79*, 223.

Vanhaudenhuyse, A., Noirhomme, Q., Tshibanda, L. J., Bruno, M. A., Boveroux, P., Schnakers, C., Soddu, A., Perlbarg, V., Ledoux, D., Brichant, J. F., Moonen, G., Maquet, P., Greicius, M. D., Laureys, S., & Boly, M. (2010). Default network connectivity reflects the level of consciousness in non-communicative brain-damaged patients. *Brain, 133*, 161–171.

Vanhaudenhuyse, A., Demertzi, A., Schabus, M., Noirhomme, Q., Bredart, S., Boly, M., Phillips, C., Soddu, A., Luxen, A., Moonen, G., & Laureys, S. (2011). Two distinct neuronal networks mediate the awareness of environment and of self. *Journal of Cognitive Neuroscience, 23*, 570–578.

Zeman, A. (2001). Consciousness. *Brain, 124*, 1263–1289.

Mind Reading, Lie Detection, and Privacy

42

Adina L. Roskies

Contents

Abstract

Neuroimaging techniques can generate a variety of kinds of personal information. This chapter focuses on the potential for neuroimaging to threaten privacy by revealing mental content and discusses the scientific and ethical issues that should be considered in a neuroethical analysis of neuroimaging that may infringe on privacy. Here these considerations are illustrated with a discussion of the scientific and ethical issues that arise when trying to use neuroimaging technologies for lie detection in real-world applications. Although current methods do not significantly threaten mental privacy, it is possible that privacy rights could be infringed with further developments in neuroimaging. However, this area is highly undertheorized; more work on the foundations of the right to privacy is needed.

A.L. Roskies
Department of Philosophy, Dartmouth College, Hanover, NH, USA
e-mail: adina.roskies@dartmouth.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_123,

Introduction

Neuroimaging techniques have the potential to generate personal information of various sorts. Already they can provide clinicians with important information about an individual's health. For example, brain scans have already been shown to be useful for diagnosing current and predicting future disease states, such as dementia (Teipel et al. 2013), and they are becoming increasingly diagnostic for certain kinds of psychiatric disorders (de Oliveira-Souza et al. 2008; Du et al. 2012; Greicius 2008; Rametti et al. 2011). In addition, scanning healthy subjects for cognitive research studies occasionally results in incidental findings with medical implications, such as the presence of a brain tumor or aneurism (Illes et al. 2006; Scott et al. 2012). Personal information regarding health could be useful to a variety of parties, such as insurance companies and employers. It goes without saying that the interests of some parties may not align with those of the subject, and thus the question of privacy arises.

The potential uses of neuroimaging information extend far beyond the medical realm. Neuroimaging can also generate a wealth of information about a person's behavior, proclivities, and experiences. In some cases, this kind of information can be obtained in seemingly neutral tasks and thus could be obtained without the subject's consent. There is evidence, for example, that neuroimaging can allow inferences about certain character traits (Blackford et al. 2011; Carre et al. 2013; Laricchiuta et al. 2013; Lemogne et al. 2011; Van Schuerbeek et al. 2011); relevance to employment decisions and law enforcement is obvious. The potential to predict future dangerousness or recidivism with the help of brain data is clearly relevant to the legal system and to law enforcement (Aharoni et al. 2013; Clark et al. 2014; Nadelhoffer et al. 2010). As we get more data and develop more sophisticated classifiers, the predictive power of neuroimaging is bound to increase. Mounting evidence suggests that some standard paradigms provide information about implicit attitudes and biases, such as racial or gender biases (Azevedo et al. 2012; Bruneau and Saxe 2010; Knutson et al. 2007; Krill and Platek 2009; Van Bavel et al. 2008). The obvious application of neuroimaging technology to lie detection could have significant influence in legal proceedings, in employment, and in other types of social settings as well (Hakun et al. 2009; Kozel et al. 2005; McCabe et al. 2011; Monteleone et al. 2009; Schauer 2010; Spence 2008; Wolpe et al. 2005). Finally, some recent studies have demonstrated the ability of fMRI to discern certain aspects of mental content, raising the specter that imaging can read minds (Chang et al. 2011; Haynes 2009; Mitchell et al. 2008; Nishimoto et al. 2011; Shinkareva et al. 2011; Soon et al. 2008).

While the science behind this is fascinating, the results may be disturbing. The growing ability of neuroimaging to produce information about mental content has led to the worry that neuroimaging threatens a boundary that has been inviolate since the dawn of humanity: the privacy of the contents of one's own mind. We are accustomed to thinking that the contents of a person's mind are directly accessible only to himself; indirect access may be granted to others through voluntary acts, such as speech, gestures, or actions. While this stark picture is not strictly true, as involuntary signals such as facial expressions can also provide a window into the contents of a person's mind, it is close enough to true that the

prospect of accessing mental content through technology could be a game changer. The ability to obtain information about mental content threatens an intuitive right to privacy. How real and pressing this threat is and what should be done about it are important neuroethical questions. This chapter addresses the question of how to approach a neuroethical analysis of neuroimaging for brain reading or mind reading purposes. Doing so requires an evaluation of the science itself, in addition to consideration of the ethical issues. Studies of neuroimaging of deception are used as an illustration.

Good Science for Ethical Analysis

It is important to realistically assess what neuroimaging techniques can tell us when addressing the ethical issues of how neuroimaging data could, should, or should not be used, so is essential to understand the science behind the technique that generates the data in question. Thus, a neuroethical analysis of a potential neuroimaging threat to privacy involves two different kinds of investigations. One is a scientific assessment of precisely what kind of information can be revealed by the neuroimaging method and what kinds of inferences can legitimately be drawn from such information. It is important to understand the nature of the information, the range of possible interpretations of that information, the extent to which it can be reasonably generalized, and with what probability of error. The second aspect involves an ethical analysis of the proposed use of that information. This includes an inquiry into the ethics of obtaining that information and the ethics of its use, including whether and to what extent people have a right to keep that information private, and the circumstances under which other parties may be justified in procuring it. While it is possible to contemplate these ethical issues in a vacuum, it is not ideal, and our analysis may be led astray. Without an understanding of the nature and quality of the information at issue, the ethical analysis may fail to engage with the relevant scenario. Thus, the precise limitations of a technique may themselves have ethically relevant consequences.

Neuroscientific Analysis

Frequently, writers concerned with the ethical and legal implications of brain imaging tend to approach the science from one of two extremes. Some take an uncritical stance toward the ability of imaging to yield relevant information, quickly generalizing from crude studies or proof of principle to embracing the proposition that imaging can reveal precise mental content. Others focus so much on the limitations of the science that they fail to recognize areas in which the scientifically feasible yet imperfect study might be practically relevant. In discussions about the potential of fMRI to detect lies and/or deception, for example, one can find both extremes. A clear assessment of the ability of the techniques to reveal mental content in particular cases is an important element of the reasoning regarding the legitimacy or illegitimacy of its use.

Internal Validity

The internal validity of a study is the degree to which a conclusion based on the data is warranted. Because of the nature of the fMRI measurement and the amount of statistical analysis required to reach conclusions, there are many ways in which neuroimaging data can be misinterpreted or misanalyzed (see, e.g., the chapter on interpreting neuroimages in this volume). One worry is that if such errors are rampant in the field, then the body of literature on which applications of imaging rests may be unsound. Despite the attention that criticisms of neuroimaging methodology may have garnered (see, e.g., Vul et al. 2009), many studies are not flawed in these ways. Moreover, as techniques become more sophisticated and practitioners more aware of potential pitfalls, the quality evidence base for neuroimaging and the ability to identify which studies are part of this quality base will improve. Thus, no sweeping indictment of imaging is reasonable. However, even when brain imaging studies meet rigorous scientific criteria for basic research, the results of these studies may not be easily applied to real-world situations.

External and Ecological Validity

External validity concerns the extent to which scientific results can be generalized or applied to non-laboratory situations. Ecological validity concerns the extent to which the behavior studied in the laboratory mirrors the behavior of interest in the natural world. Because the interpretation of neuroimaging results in part depends upon comparing them to results of other imaging studies, or to generalizations based on other studies, the external validity of those studies is essential to assess. And because we often are interested in a naturally occurring phenomenon and try to discover its neural basis in the lab, it is also important to assess ecological validity. In what follows, the study of interest is termed the target study and the background data as comparison studies. The assessment of the external validity of neuroimaging studies is crucial in determining whether inferences based on brain imaging data are warranted, and assessment of ecological validity is important for determining whether such research can be applied in real-world settings. A number of the most significant factors to consider when making such assessments are discussed below.

What Is Being Measured in a Study?

Neuroimaging paradigms are not always what they seem, and researchers sometimes set out to study a phenomenon but end up designing a paradigm that unwittingly probes something else. In all cases it is essential to determine what is being measured in a study. Is it what the study reports it measures? Secondly, is what is being measured in the target study the same thing that is being measured in the intended comparison studies? If not, how are they related? Establishing this is important in order to assess the degree to which the target study can be interpreted in light of the intended background.

What Is the Population Studied? Is There Reason to Think It Differs from the General Population?

There is significant variability in neuroimaging results across different populations. Most basic research is carried out on normal, healthy college-age populations. One should always be aware of whether the target study involves a different population. Is there reason to think there will be systematic differences between responses of these populations? The answer may be no if the target study involves normal adults, but we know that there are significant age-related changes in brains of juveniles and the elderly. There may also be relevant differences among socioeconomic extremes. Moreover, some target studies may focus on particular populations that we have reason to think may have substantial neural differences from the norm, such as drug addicts, violent offenders, or people with psychiatric dysfunction (Brower and Price 2001). It has been reported that a large percentage of incarcerated populations have sustained damage to frontal areas and thus have brains that are significantly different from normal on a variety of measures (Brower and Price 2001). One has to be very careful about making inferences about the significance of deviations from normal activation patterns in such groups.

What Is the Error of the Measurement?

If using brain imaging for diagnostic or predictive purposes, it is important to understand error rates. However, error rates in many studies are not well defined and pose statistical problems discussed further below. In addition, using imaging for diagnostics requires that one have statistics about the baseline occurrence of a phenomenon of interest. Where this information is not available, true error rates may not be able to be assessed.

Are There Statistical Limits to the Application?

Brain imaging is noisy: many trials are averaged together in order to reduce noise so signal can be detected. Often trials are averaged among subjects, confounding differences in brain structure and functional anatomy with results. The poor signal to noise of fMRI makes it very difficult to apply in paradigms where repetition is difficult or impossible or where it alters the nature of the task.

Ethical Analysis

A number of ethical and legal considerations also impact whether and to what extent we should use the results of neuroimaging to reveal mental content. These considerations regard the manner in which that information is obtained and whether obtaining it infringes rights that are either fundamental or otherwise enshrined in the law. Below I discuss a number of issues that should be considered in an ethical analysis of neuroimaging.

Is the Data Obtained Ethically?

Harm

Obvious ethical questions concern the procedures of obtaining neuroimaging data. A first question is whether the process of obtaining the data results in physical harm to the subject. For example, some imaging or related techniques, such as those employing radiation (such as PET) or disrupting neural firing (such as TMS), may potentially, under some regimes, harm the subject. These harms in many cases would rule out the use of the technique and in other cases must be weighed against potential benefits. However, there is no evidence that techniques most frequently employed for brain reading, fMRI and EEG/ERP, pose any health threat.

Consent

A second question concerns consent. The ethical questions about using neuroimaging-derived information vary depending upon whether the information is obtained with or without consent. If a subject freely consents to a study, some worries are allayed. However, since some information can be gleaned without a subject's knowledge, even subjects consenting to be scanned may not have consented were they fully informed of the type of information to be gleaned. The potential of fMRI to yield information about some psychological aspect other than the one being overtly probed raises ethical questions about the effectiveness or legitimacy of standard consent practices.[1] Information of that sort should perhaps be treated as data obtained without consent, and when prior consent for this type of information is not possible, perhaps in such cases post hoc consent should be sought before the data is used.

In some potential uses of neuroimaging, such as in legal or employment contexts, consent may be coerced. In general, however, coercion is unlikely to be effective, because fMRI requires cooperation by the subject in several ways: subjects must remain very still in order to get interpretable data; slight head motion or even movement of the jaw or tongue can make scan data unusable. Secondly, most fMRI studies require that subjects perform a task, which also requires cooperation. Because of the practical necessity for cooperation, we are likely to see brain scans introduced much more in situations in which they could be potentially helpful to the subject. Thus, they are more likely to be introduced as evidence in exculpatory contexts in criminal cases or as evidence of truth telling rather than lying.

Of course, if a technique is to be used without consent, other questions arise. First, is the technique effective in a non-consenting subject? Second, can results be effectively subverted with countermeasures? Because neuroimaging is currently well-nigh impossible on a normal subject that does not consent at least to being the subject of an experiment, a variety of potential objections do not arise. One possible scenario, however, is that as we learn more about what can be gleaned by

[1] It should be noted that this is already possible to achieve with behavioral tests. Moreover, it is not clear that this information tells us anything about how a person will act in a certain situation. The biases, even if real, could be counteracted by other processes and never affect decision or action.

baseline or resting-state fMRI, certain things, such as disease states, may be discernable from scans that do not require active subject participation. Another notable exception arises when the subject is unable to consent, such as when the subject is in a state of diminished consciousness (see ▶ Chap. 41, "Detecting Levels of Consciousness"). In such cases, consent should be sought from a proxy.

Are There Alternative Methods for Getting the Same Data?

Often, if you want to know what a person is thinking, the best way to find out is to ask him. In many cases, this is preferable to trying to determine an answer through imaging. However, there are situations in which a consenting subject could not otherwise provide the sought-after information. In some cases, this information is not available to consciousness. Studies of bias using a number of behavioral paradigms probe such information. In other cases, such as in lie detection, subjects have an issue with credibility. Because the credibility of the subject is at issue, voluntary disclosure does not substitute for a measure that has the ability to detect deception. One then has to ask whether the neuroimaging ability to detect deception in the kind of circumstances at issue is sufficiently good for the purpose at hand. The scientific analysis is crucial here. Moreover, if it is more accurate than other available techniques, are there nonetheless reasons to prefer those other techniques? In legal settings, there is an abiding sentiment that juries are a more appropriate lie detection method than a scientific/mechanical one, even though the justification for this is rarely made explicit, and it is not clear that there is a defensible one. How good would a scientific technique have to be to substitute for a jury decision? Do the consequences in terms of perceived legitimacy, potential misuse, etc., require only that the technique be more reliable than jury decisions or much more so?

Is There a Right Being Violated?

Perhaps the central ethical issue raised by neuroimaging is whether it infringes upon a right. The standard worry is that brain imaging violates a right we have to the privacy of the contents of our thoughts. Until now, no techniques have been developed that accurately reveal mental content without behavior. Now the prospect exists for obtaining evidence for mental content without reference to behavior, whether voluntary or involuntary.

Privacy issues arise when someone other than the subject desires to procure information that the subject may want to keep private. Then one must answer the difficult questions of whether the person has a right to keep that information private and, if so, whether there is justification for infringing that right.

Is There a Right to Privacy?

Although a right to mental privacy is often taken to be intuitive, it is not clear whether there are deep principled reasons that undergird such a right or whether such a right is assumed simply because the mental content is by nature private, and breaching that privacy usually involves coercive techniques. The foundations for

such a right are not clearly articulated, though it is argued that privacy is a necessary protection for freedom. One's view of whether one has a right to the privacy of the contents of one's thoughts and the scope of such a right will likely depend at least in part on other ethical commitments. For example, utilitarians will probably have very different views about privacy rights than those with a more deontological bent.

As mentioned, although intuitively there is value to the first-person access one has to the content of one's own thoughts, the nature of that value is obscure. Little has been written philosophically defending the value of mental privacy, but as one aspect of what is sometimes called "cognitive liberty," it is frequently taken to be a fundamental right (see, for instance, http://www.cognitiveliberty.org/; Farahany 2012; Warren and Brandeis 1890). Alternatively, as Warren and Brandeis theorized, mental privacy can be thought of as an extension of common law property rights. They analogized thoughts to a creation of the thinker and recognized a right in the common law that gives to each author sole rights to his creation (e.g., thought) until he voluntarily communicates it to the public. "The common law secures to each individual the right of determining, ordinarily, to what extent his thoughts, sentiments, and emotions shall be communicated to others. Under our system of government, he can never be compelled to express them (except when upon the witness stand); and even if he has chosen to give them expression, he generally retains the power to fix the limits of the publicity which shall be given them." Warren and Brandeis have perhaps best articulated this implicit broad right to privacy for mental content and argued that mental content, regardless of how expressed, should be afforded the same protection. They continue, ". . .If, then, the decisions indicate a general right to privacy for thoughts, emotions, and sensations, these should receive the same protection, whether expressed in writing, or in conduct, in conversation, in attitudes, or in facial expression." It is but a small step to extend the same protections to mental contents that remain unexpressed.

Warren and Brandeis' argument is influential and perhaps compelling, but it is not law. There is no express right to privacy in the Constitution. The Bill of Rights, however, protects specific aspects of privacy, although in a piecemeal and less explicit way, protecting what Supreme Court Justice Louis Brandeis dubbed as a "right to be left alone," in Olmstead v. United States, 277 U.S. 438 (1928). The First Amendment protects certain freedoms, such as the freedom of religion, of the press, and of expression, prohibiting federal laws restricting these freedoms. The Amendment has been read broadly as protecting the privacy of beliefs, but whether these protections would be extended to mental states discernable by neuroimaging technologies has not been explored in the case law. At a theoretical level, first amendment rights may not appear to be directly applicable to neuroimaging scenarios, for they concern freedom of expression. Nothing about neuroimaging, which is a measuring technique, interferes with expression, nor does imaging involve any sort of manipulation or restriction of thought content. However, the courts have come down strongly not only on measures that restrict expression but also on those that can have a "chilling effect" on expression.

Because it is very likely that if neuroimaging could be used to discern the contents of thoughts against a person's will, it would have a chilling effect not only on expression but also on the source of expression, and thus it would impact the freedom people have even to entertain those thoughts. It is here that First Amendment protections would most likely come into play.

While the First Amendment protects expression and belief, the Fourth Amendment protects the right of citizens to be secure in their persons and effects against unreasonable searches and seizures. It is no stretch to imagine that brain scanning without consent, for instance, would constitute an unreasonable search, and even potentially that mental content revealed could be considered property, as Warren and Brandeis have argued. The Fifth Amendment protects persons against self-incrimination, preventing the state from using a person's own knowledge against himself. The uncertain status of imaging as physical or testimonial evidence will be crucial here, for if it is seen as physical evidence, the protection does not apply. The Fourth and Fifth Amendments thus have the potential to provide for a privacy of the mental insofar as mental states are aspects of persons abut which expectations of privacy are reasonable or are testimonial and can be incriminating.

These rather circumscribed areas of privacy rights are arguably significantly broadened by the Ninth Amendment, which states that the people retain rights not necessarily enumerated in the Bill of Rights. However, what those rights are, if there are any, is highly disputed. Finally, the Fourteenth Amendment extends the protections in the Bill of Rights against infringements by the states. The Due Process clause of that amendment mandates that the states do not deprive citizens of life, liberty, or property without due process of law. The liberty interests thus protected have been interpreted by the courts to include a variety of substantive freedoms such as the right of contract and right of privacy in one's home and relations, and in the last decades substantive due process has been extended to many other aspects of personal life. Sometimes also glossed as the "right to be left alone," this aspect of the Fourteenth Amendment may arguably ground a more extensive domain of privacy than any previously discussed.

Is It Absolute and Inviolable or Is It a Presumption That Is Weighed Against Other Considerations?

In the law, these privacy rights, and by extension privacy afforded to mental content, are not inviolable, but are weighed against the needs of the state. Moreover, while legal strictures against infringement against freedom of expression are stringent, requiring strict scrutiny, those regarding private property are less so. Thus, even though the law recognizes that aspects of privacy deserve to be protected, those protections are not absolute and privacy can be infringed given sufficient reason. Given the recent revelations about the scope of government surveillance on citizens and foreign nationals, it has become a matter of much debate how we should weigh privacy rights against other rights and the needs of the state. Clearer articulation is needed of the importance of privacy, both as an end in itself and, instrumentally, for the preservation of other values.

Application: Detection of Deception

To illustrate the above concerns, one prominent potential use of neuroimaging of mental content is discussed: whether laboratory studies of lie detection ground the applicability of fMRI lie detection for real-world applications.

Scientific Analysis

External and Ecological Validity

There is no doubt that neuroimaging can provide some information relevant to lie detection or detection of deception (Christ et al. 2009; Hakun et al. 2009; Kozel et al. 2005). Numerous studies demonstrate that a constellation of brain areas is more active during deception than during non-deceptive responding in these paradigms. Despite this, recent meta-analyses show that although deception reliably activates a number of brain areas relative to truth telling, there is considerable variability across studies, and no regions were found to be activated in all studies (Christ et al. 2009). There is thus no specific fMRI brain signature of deception. Furthermore, there is reason to question whether these methods probe mental content at all: these paradigms may not isolate neural changes reflective of the phenomenon of lying itself or even of the intent to deceive and may instead reflect brain activation that merely co-occurs regularly with deception (such as, e.g., emotional arousal, salience, etc.). For discussion of a number of experimental confounds, see Farah et al. (2014). While mere correlation may suffice for some purposes, for others – such as determining whether fMRI indications of deception is more like physical evidence or testimony, an important question regarding the constitutionality of fMRI lie detection in court – it cannot. Finally, there are significant reasons to question the applicability of fMRI lie detection to practical situations, such as legal contexts (Langleben and Moriarty 2013; Wagner 2010; Wolpe et al. 2005).

Perhaps the main problems with experiments on lie detection have to do with whether the paradigms used in extant studies really probe or measure the phenomenon they purport to measure: lying or deception. Many of the studies employ paradigms in which subjects are instructed to lie. In some, they are instructed to lie about a particular issue (i.e., lie when queried one of two cards they hold (Davatzikos et al. 2005; Langleben et al. 2005)); in others they are told to lie about a number they picked (Hakun et al. 2008) or are cued to lie or tell the truth about a question (Ganis et al. 2003); in still others they are instructed to perform an act and then lie about it (i.e., “steal” a ring or watch, then lie about having done it) (Kozel et al. 2005). However, in all these cases, it is questionable whether the subjects are really engaging in deception or are better thought of as complying with instruction. This illustrates the difficulty in designing studies that incorporate ecologically valid deception tasks. Perhaps the only extant study not subject to this worry is one in which subjects are not instructed to deceive and instead are merely given the

opportunity to do so, and when they do, they do so out of their own motivation believing that they are gaming the system. In this study by Greene and colleagues, subjects are told to report the outcomes of blind coin tosses, in a money game. Deceptive subjects are identified on the basis of improbable winnings (Greene and Paxton 2009). This is likely the most ecologically valid of the deception studies; its limitation is that the outcomes of individual trials are unknown, so individual subject behavior can only be determined statistically, across many trials. While there are differences in overall brain activation patterns between subjects who lie more often and those that lie less, these differences cannot be determined on a trial by trial basis, and they may be due to factors that correlate with the tendency to deceive, such as, for instance, particular character traits, but not with deception itself.

A more subtle issue about what is being measured is illustrated by some lie detection paradigms. In some studies, the conditions in which subjects are instructed to lie tend to occur infrequently and involve the detection of a target. For instance, an instruction to "lie when the Queen of Hearts comes up" may seem like a lying task, but it may better be described as a detection task for the Queen of Hearts, rather than a lying task (Emilio Kanwisher's contribution in Bizzi et al. 2009). Indeed, the dynamical brain signatures of such tasks look much like oddball detection paradigms that do not involve lying (Rosenfeld et al. 2008). Thus, a study may masquerade as a lie detection task but may in fact measure a quite different psychological phenomenon.

Finally, a variety of contextual factors differ between laboratory settings and real-world settings in which lie detection technologies would be most useful. The stakes in real-world settings are much higher than in the lab, and this can generate emotional reactions not present in experimental settings. Many of the brain areas involved in emotional response and regulation are precisely the ones that display differences in lie detection imaging results, leading to the possibility that using fMRI lie detection paradigms (or, for that matter, other physiological measures) in realistic situations will generate a significant number of false positives.

Population Differences

Another reason to be suspicious of the external validity of lie detection paradigms is that the population that these studies draw upon differs significantly from the populations upon which lie detection devices are most likely to be used. As noted before, fMRI research studies tend to use college students; lie detection technologies are most likely to be used in forensic contexts or in personnel screening. The demographics of these populations are much different. Whether results from college students would be representative of or extendable to these diverse populations is unknown.

Errors and Statistics

There are also significant statistical limitations to using the data from lie detection experiments in real-world situations. Applications of lie detection methods generally aim to determine whether the response of an individual to a particular question or about a particular issue is a lie. Thus, the utility of these techniques depends upon

being able to distinguish lies from sincere responses for individuals on individual trials or items. However, the vast majority of lie detection studies that could be used as comparison tasks rely upon data averaged across multiple subjects, each of whom had undergone multiple trials. Moreover, most studies report areas consistently activated across subjects, but do not report individual subject data. This provides a prima facie barrier for using this information for identifying lies in individuals or on individual trials. The few studies that attempt to determine whether individuals are lying on specific trials (Davatzikos et al. 2005; Langleben et al. 2005) report lower accuracy than group studies and suffer from experimental design confounds making these figures suspect (Farah et al. 2014).

Finally, an underappreciated factor may be a significant barrier to understanding the diagnostic power of neuroimaging for lie detection. Although several companies have already begun to offer lie detection services to consumers, touting their accuracy, it is not possible to estimate the actual diagnostic power of the technique without an understanding of how prevalent lying or deception is in the populations of interest. For example, No Lie MRI advertises that their technique is over 90 % accurate and estimates that in the future accuracy may reach 99 % (http://www.noliemri.com). But these predictions are misleading. Knowledge of base rates and the specificity and sensitivity (rates of false positives and false negatives) of methods is needed to provide a full picture of how good a technique is. Few studies report the sensitivity and specificity of the methods, and when they do, they are relative to the laboratory test. Whether error rates associated with a diagnostic method is acceptable will depend upon the base rate of lying, and no measure of base rates for lying is available, nor is it clear how it could ever be. Furthermore, even if one could determine base rates for lying, what would be the relevant population and context for which the base rate should be determined? Would it be the general rate of lying in the public? The rate of lying for people charged for a crime and undergoing lie detection? The rate of lying in the prison population?

Without knowledge of base rates and error rates, the predictive accuracy of these techniques cannot be determined. Indeed, in two recent court cases (US v Semrau 2010; Wilson v. Corestaff Services 2010), the courts decided against admitting neuroimaging evidence for lie detection. Because of unknown error rates and worries about the ecological validity of the laboratory tests, neuroimaging for lie detection was deemed not to pass the Daubert or Frye standards for scientific evidence.

Ethical Analysis

Harm and Consent

Typical neuroimaging paradigms are not harmful and require the cooperation of the subject. However, to the extent that they can be used to discriminate familiar from unfamiliar information with passive viewing, they could potentially be used without the subject's knowing what they are in the scanner for. This may be considered a use without consent, which raises an ethical flag.

Privacy

Does the use of neuroimaging for lie detection infringe on privacy rights? Certainly neuroimaging can provide information in which people could have a clear privacy interest, such as in health-related information mentioned in the introduction. These concerns are real, but they are not significantly different from the concerns raised by genetic and other personal information, and there are already models in the health sciences for dealing with these issues through HIPPA and related doctrines. Recently, some heightened concerns regarding anonymity have arisen with regard to genetics because it has been shown that in some circumstances enough information can be gleaned from anonymous genetic data to determine the person from whom it came (Gymrek et al. 2013). However, thus far there is little reason to think that brain imaging data will present the same kind of threat of carrying covert identifying information that genetics does: although our brains are certainly unique, there are no analogs here to heritability and surname inheritance to aid in identification, and we lack information about brain function at a fine-grained enough level to ground identifying inferences between brain data and personal characteristics.

What makes neuroimaging of deception potentially threatening is not just the dissemination of personal information; it is the specter of abolishing the privacy of the mental. With regard to First Amendment protections, one should ask whether lie detection poses a threat to freedom of expression or belief. Since it does not restrict expression or prohibit any beliefs, it is unlikely that imaging will be prohibited for that reason; it is also not likely that the prospect of lie detection would be considered to have a chilling effect on protected speech. Some paradigms, such as the guilty knowledge paradigm (GKT), which aims to distinguish brain activations to familiar and unfamiliar stimuli, may abridge privacy rights. Fourth Amendment protections could be invoked to prevent the state from imaging a person without their consent, as this may be thought to constitute unreasonable search of a person's person or property. However, with current technology, it is very difficult to image a noncooperative subject, so it is unlikely that this defense will prove necessary in the short term. Fifth Amendment protections might also be invoked: there is an interesting theoretical question about whether neuroimaging data is like physical evidence (such as fingerprints and blood alcohol level), and thus not protected by the Fifth Amendment, or like testimony, and thus should enjoy Fifth Amendment protections (see, e.g., Farahany 2012; Fox 2008; Gymrek et al. 2013; Holley 2009; Hurd 2012; Stoller and Wolpe 2007). However, the scientific limitations of lie detection technologies and the difficulty of applying them in real-world situations make it unlikely that any of these protections will need to be invoked any time soon, because these techniques are not yet admissible in court. Thus, at least for the foreseeable future, it seems that privacy concerns are not going to be the ones that limit the use of neuroimaging; concerns about accuracy and applicability will be.

One area in which lie detection seems promising, is for use in exculpation. First, if imaging is used in order to exculpate, coercion and privacy concerns are further negated. More importantly, although the limited accuracy of lie detection and questions about its diagnostic power are reason to preclude its use in inculpating contexts, the statistics are in its favor for providing some support for indicating

absence of lying. Moreover, it is worth asking about the relative costs of error: given our presumption of innocence and the violence to our values done by convicting the innocent, a false negative (finding a guilty person innocent) is arguably more tolerable than a false positive (finding an innocent person guilty). Thus, it may be reasonable to hold imaging to different standards of accuracy depending on the context (Schauer 2010). Thus far, it seems that courts are unwilling to admit lie detection evidence even in exculpatory contexts, perhaps because of the prospect of the slippery slope that may lead to them being admitted in inculpatory contexts, where the statistics are not in their favor.

Conclusion and Future Directions: Mind Reading More Broadly

For illustration this chapter focused specifically on uses of lie detection. This is not ideal as a paradigm for privacy of mental content, since it is unclear whether lie detection methods even probe the content of thoughts. However, it is the only arena to date that is arguably related to mind reading in which significant work has been done geared toward potential use in real-world applications. Even so, for the numerous reasons here discussed, current technology is not sufficiently well developed to use lie detection for incrimination or even for screening. In other kinds of work, privacy issues may be raised much more clearly. There are increasingly effective attempts in the laboratory for identifying elements of mental content, such as what object a person is thinking of or even what visual scene they are experiencing. While the ability to reconstruct visual perceptual experiences from data garnered while they are being experienced is fairly impressive (Kay et al. 2008; Naselaris et al. 2009; Nishimoto et al. 2011), these methods do not do particularly well with memory reconstruction or identification of abstract thought. For the most part, current methods only allow inferences about rather coarse-grained content, and we are far from being able to reconstruct determinate propositional content (see Roskies forthcoming). Whether thought content, in all its richness and subtlety, could ever be a realistic target for imaging remains to be seen. Moreover, the successes thus far have been successes with a cooperating subject. There is scant evidence that mental content in any detail can be determined without cooperation of the subject. Thus, current methods are far from allowing determinate mental content to be gleaned from a non-consenting subject (Roskies forthcoming), reducing the potential range of scenarios that are most ethically challenging. If this should change appreciably, however, we will have to revisit the privacy issues sketched above in considerably more detail.

Cross-References

► A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience
► A Duty to Remember, a Right to Forget? Memory Manipulations and the Law
► Cognitive Liberty or the International Human Right to Freedom of Thought
► Neurolaw: Introduction

References

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., et al. (2013). Neuroprediction of future rearrest. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 6223–6228.

Azevedo, R. T., Macaluso, E., Avenanti, A., Santangelo, V., Cazzato, V., & Aglioti, S. M. (2012). Their pain is not our pain: Brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Human Brain Mapping, 34*, 3168–3181.

Bizzi, E., Hyman, S. E., Raichle, M. E., Kanwisher, N., & Phelps, E. A. (2009). *Using imaging to identify deceit: Scientific and ethical questions*. Cambridge, MA: American Academy of Arts and Sciences.

Blackford, J. U., Avery, S. N., Cowan, R. L., Shelton, R. C., & Zald, D. H. (2011). Sustained amygdala response to both novel and newly familiar faces characterizes inhibited temperament. *Social Cognitive and Affective Neuroscience, 6*, 621–629.

Brower, M. C., & Price, B. H. (2001). Neuropsychiatry of frontal lobe dysfunction in violent and criminal behaviour: A critical review. *Journal of Neurology, Neurosurgery and Psychiatry, 71*, 720–726.

Bruneau, E. G., & Saxe, R. (2010). Attitudes towards the outgroup are predicted by activity in the precuneus in Arabs and Israelis. *NeuroImage, 52*, 1704–1711.

Carre, A., Gierski, F., Lemogne, C., Tran, E., Raucher-Chene, D., et al. (2013). Linear association between social anxiety symptoms and neural activations to angry faces: From subclinical to clinical levels. *Social Cognitive Affective Neuroscience*. doi: 10.1093/scan/nst061.

Chang, K.-M. K., Mitchell, T., & Just, M. A. (2011). Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *NeuroImage, 56*, 716–727.

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex, 19*, 1557–1566.

Clark, V.P., Beatty, G.K., Anderson. R.E., Kodituwakku. P., Phillips. J.P., et al. (2014). Reduced fMRI activity predicts relapse in patients recovering from stimulant dependence. *Human Brain Mapping, 35*, 414–428.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., et al. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage, 28*, 663–668.

de Oliveira-Souza, R., Hare, R. D., Bramati, I. E., Garrido, G. J., Azevedo Ignacio, F., et al. (2008). Psychopathy as a disorder of the moral brain: Fronto-temporo-limbic grey matter reductions demonstrated by voxel-based morphometry. *NeuroImage, 40*, 1202–1213.

Du, W., Calhoun, V. D., Li, H., Ma, S., Eichele, T., et al. (2012). High classification accuracy for schizophrenia with rest and task FMRI data. *Frontiers in Human Neuroscience, 6*, 145.

Farah, M.J., Hutchinson, B., Phelps, E.A., Wagner, A.D. (2014). fMRI lie detection: Scientific and societal challenges. *Nat Review Neurosciences, 15*, 123–131.

Farahany, N. (2012). Incriminating thoughts. *Stanford Law Review, 64*, 351–408.

Fox, D. (2008). Brain imaging and the bill of rights. *The American Journal of Bioethics, 8*, 34–36.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex, 13*, 830–836.

Greene, J., & Paxton, J. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *PNAS, 106*, 12506–12511.

Greicius, M. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Current Opinion in Neurology, 21*, 424–430.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science, 339*, 321–324.

Hakun, J. G., Seelig, D., Ruparel, K., Loughead, J. W., Busch, E., et al. (2008). FMRI investigation of the cognitive structure of the concealed information test. *Neurocase, 14*, 59–67.

Hakun, J. G., Ruparel, K., Seelig, D., Busch, E., Loughead, J. W., et al. (2009). Towards clinical trials of lie detection with fMRI. *Social Neuroscience, 4*, 518–527.

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences, 13*, 194–202.

Holley, B. (2009). It's all in your head: Neurotechnological lie detection and the fourth and fifth amendments. *Developments in Mental Health Law, 28*, 1–23.

Hurd, A. J. (2012). Eaching past fingertips with forensic neuroimaging—non-"Testimonial" evidence exceeding the Fifth Amendment's Grasp. *Loyola Law Review, 58*, 213–248.

Illes, J., Kirschen, M. P., Edwards, E., Stanford, L. R., Bandettini, P., et al. (2006). Ethics. Incidental findings in brain imaging research. *Science, 311*, 783–784.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*, 352–355.

Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human Brain Mapping, 28*, 915–930.

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry, 58*, 605–613.

Krill, A., & Platek, S. M. (2009). In-group and out-group membership mediates anterior cingulate activation to social exclusion. *Frontiers in Evolutionary Neuroscience, 1*, 1.

Langleben, D. D., & Moriarty, J. C. (2013). Using brain imaging for lie detection: Where science, law and research policy collide. *Psychology, Public Policy, and Law: An Official Law Review of the University of Arizona College of Law and the University of Miami School of Law, 19*, 222–234.

Langleben, D. D., Loughead, J. W., Bilker, W. B., Ruparel, K., Childress, A. R., et al. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human Brain Mapping, 26*, 262–272.

Laricchiuta, D., Petrosini, L., Piras, F., Cutuli, D., Macci, E., et al. (2013). Linking novelty seeking and harm avoidance personality traits to basal ganglia: Volumetry and mean diffusivity. *Brain Structure & Function*. doi: 10.1007/ s00429-013-0535-5.

Lemogne, C., Gorwood, P., Bergouignan, L., Pelissolo, A., Lehericy, S., & Fossati, P. (2011). Negative affectivity, self-referential processing and the cortical midline structures. *Social Cognitive and Affective Neuroscience, 6*, 426–433.

McCabe, D. P., Castel, A. D., & Rhodes, M. G. (2011). The influence of fMRI lie detection evidence on juror decision-making. *Behavioral Sciences & the Law, 29*, 566–577.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*, 1191–1195.

Monteleone, G. T., Phan, K. L., Nusbaum, H. C., Fitzgerald, D., Irick, J.-S., et al. (2009). Detection of deception using fMRI: Better than chance, but well below perfection. *Social Neuroscience, 4*, 528–538.

Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K., Mansfield, A., et al. (2010). Neuroprediction, violence, and the law: Setting the stage. *Neuroethics*, 1–33.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron, 63*, 902–915.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology, 21*, 1641–1646.

Rametti, G., Carrillo, B., Gomez-Gil, E., Junque, C., Zubiarre-Elorza, L., et al. (2011). The microstructure of white matter in male to female transsexuals before cross-sex hormonal treatment. A DTI study. *Journal of Psychiatric Research, 45*, 949–954.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology, 45*, 906–919.

Roskies, A. L. (forthcoming). Mindreading and privacy. In M. Gazzaniga (Ed.), *The cogntive neurosciences V*.

Schauer, F. (2010). Neuroscience, lie-detection, and the law. *Trends in Cognitive Sciences, 14*, 101–103.

Scott, N. A., Murphy, T. H., & Illes, J. (2012). Incidental findings in neuroimaging research: A framework for anticipating the next frontier. *Journal of Empirical Research on Human Research Ethics: JERHRE, 7*, 53–57.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *NeuroImage, 54*, 2418–2425.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience, 11*, 543–545.

Spence, S. A. (2008). Playing Devil's advocate: The case against fMRI lie detection. *Legal and Criminological Psychology, 13*, 11–25.

Stoller, S. E., & Wolpe, P. R. (2007). Emerging neurotechnologies for Lie detection and the fifth amendment. *American Journal of Law & Medicine, 33*, 359–375.

Teipel, S. J., Grothe, M., Lista, S., Toschi, N., Garaci, F. G., & Hampel, H. (2013). Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease. *The Medical Clinics of North America, 97*, 399–424.

US v Semrau. (2010). 643 F. 3d 510, decided Sept. 7, 2012

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological Science, 19*, 1131–1139.

Van Schuerbeek, P., Baeken, C., De Raedt, R., De Mey, J., & Luypaert, R. (2011). Individual differences in local gray and white matter volumes reflect differences in temperament and character: A voxel-based morphometry study in healthy young females. *Brain Research, 1371*, 32–42.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Wagner AD. (2010). Can neuroscience identify lies? In Gazzaniga, Michael S., (Eds.), *A judges guide to neuroscience: A concise introduction* (pp. 13–25). SAGE Center, UC Santa, Barbara.

Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard Law Review, 4*, 193.

Wilson v Corestaff (2010) 28 Misc. 3d 425. (May 14, 2010)

Wolpe, P. R., Foster, K. R., & Langleben, D. D. (2005). Emerging neurotechnologies for lie-detection: Promises and perils. *The American Journal of Bioethics, 5*, 39.

Section IX

Ethical Implications of Brain–Computer Interfacing

Ethical Implications of Brain–Computer Interfacing

43

Jens Clausen

Contents

Abstract

Brain–computer interfaces enable direct brain-to-computer communication circumventing peripheral nerves and muscles. These devices offer promising clinical applications for a broad range of patients. This paper presents a brief overview of different BCI applications and ethical issues raised by these devices, and gives an introduction to this section containing four chapters focusing on the ethical aspects of BCIs for communication, for motor prosthetics, in ADHD patients, and in psychopathy, respectively.

Introduction

Connecting brains to computers usually relies on peripheral nerves and muscles as well as keyboards or other input devices like a computer mouse. Direct brain-to-computer communication bypassing peripheral nerves and muscles—for a long time a topic of the science fiction genre—is now possible through brain–computer interfaces (BCI). Miniaturization of microelectronic components and continuous progress in brain research together with an ever better understanding of the human

J. Clausen
Institute for Ethics and History of Medicine, University of Tübingen, Tübingen, Germany
e-mail: jens.clausen@uni-tuebingen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_40,

brain enables unprecedented direct brain-to-computer communication. The papers of this section briefly introduce different BCI approaches and discuss key ethical questions related to this technology.

Brain–Computer Interfaces

Usually, a BCI records neural signals from the brain or parts of it and uses these signals for controlling an external device. In general, three parts constitute a BCI: First, an internal interface enables signal acquisition, second, processing of these signals is carried out through algorithms in a central computing unit, and third, an external interface ensures contact to the surroundings via a specific output device (Wolpaw et al. 2002; Clausen 2008; van Gerven et al. 2009). Depending on the exact shape of these components, BCIs differ significantly (Berger et al. 2007) and may be used in a vast variety of diverse applications (Mak and Wolpaw 2009). These include assistive technologies like spelling devices for the severely paralyzed (Birbaumer et al. 1999; Birbaumer and Cohen 2007), motor prostheses for paraplegics (Hochberg et al. 2012; Collinger et al. 2013), neurorehabilitation for stroke patients (Buch et al. 2008), and a BCI-controlled wheelchair (Leeb et al. 2007; Galán et al. 2008). BCIs might also be used outside the clinic. While developed for creative expression in paralyzed patients (Munssinger et al. 2010; Zickler et al. 2013), BCI brain painting is already used by artists (pingo-ergo-sum.com). Controlling computer games also seems a possible application of BCIs (Liao et al. 2012). This section, however, focuses on clinical BCIs and their ethical implications.

Ethical Implications

While BCIs promise various benefits for different groups of patients, these devices also need an ethical examination. The ethical discussion started about 10 years ago and addresses a broad range of issues including quality of life, responsibility and liability, the notion of autonomy in using such devices, personal identity, and questions of enhancement when used beyond the realm of therapy (Clausen 2006, 2008, 2009, 2011; Johansson 2008; Lucivero and Tamburrini 2008; Haselager et al. 2009; Tamburrini 2009; Vlek et al. 2012).

A recent special issue in *Neuroethics* (Clausen 2013) and a report from the Nuffield Council on Bioethics (Baldwin et al. 2013) present a comprehensive overview on neurotechnologies including stimulating devices and their ethical implications. This section of the Handbook of Neuroethics addresses the ethical implications of BCIs for communication, motor control, BCI in attention deficit hyperactivity disorder (ADHD), and also BCI approaches in Psychopathy.

Surjo Soekadar and Niels Birbaumer focus their paper on brain–machine interface-based communication especially in complete paralysis like the locked-in state. They present an introduction to the different technological approaches of brain–machine or brain–computer interfaces followed by current technical

challenges in the use of BCI for paralyzed patients like the high rate of false-positive and false-negative classifications and the missing evidence for BCI control in the complete locked-in state (where no residual voluntary motor control is left). This might suggest a fundamental problem of learning and attention. While fragmentation of sleep challenges BCI performance especially with respect to hitting the right time for recordings, the hypothesis of the extinction of goal-directed thinking due to the lack of reinforcing feedback presents a more fundamental challenge. If this hypothesis is true, it might be difficult if not impossible to establish BCI-based communication with patients in the complete locked-in state. However, implementing classical semantic conditioning into a metabolic BCI based on near-infrared spectroscopy might be a promising alternative. Soekadar and Birbaumer identify and discuss ethical implications of BCI-based communication in paralysis with respect to questions of informed consent, advance directives, and quality of life. They point to evidence that physicians and relatives usually underestimate these patients' quality of life. This might also be true for the patients themselves when in a premorbid stage anticipating the quality of life in a locked-in state, which might challenge the validity of end-of-life decisions in such advanced directives. Against the backdrop of high rates of misdiagnosis in disorders of consciousness, the authors recommend EEG monitoring and the use of BCI systems in behavioral unresponsive patients for detecting possible remaining awareness.

Donatella Mattia and Guglielmo Tamburrini focus their investigation on the ethical implications of BCI systems for motor control. They address the ethical challenges of these devices around five conceptual frameworks: first informed consent and respect for persons, second beneficence, justice, and autonomy, third responsibility and liability, fourth liberty, and fifth responsible communication of BCI research. The authors concede that these ethical issues are not unique questions raised by BCIs but are well known from other fields of neuroethics and biomedical ethics as well. However, Mattia and Tamburrini identify three aspects in the field of BCI research, which recommend a closer look at the mentioned ethical issues. *Brain–computer mutual adaptation* points to the fact that BCI operation involves two adaptive controlling units: the central nervous system and a computer algorithm. *Machine intelligence* describes the need for autonomous intelligent action of BCI-controlled devices at least to some degree. The third aspect emphasizes the restricted class of *potential users* of BCI devices for motor control. Persons affected by severe motor impairments are most likely to benefit from this kind of technology, while those who are less affected and kept the ability to some voluntary movements might prefer alternative assistive technology.

Imre Bard and Ilina Singh introduce attention deficit hyperactivity disorder (ADHD) and BCI-neurofeedback followed by a summary of today's knowledge on BCI-neurofeedback in ADHD. In the current state of limited knowledge and contradictory results, they identify safety and efficacy as key elements for the ethical discussion in this field of BCI technology. Against this backdrop, they address questions of regulation and commercialization, responsible communication and neuroliteracy, identity, and the use of BCI for performance enhancement. Bard and Singh emphasize two different but interrelated challenges:

commercializing BCI-neurofeedback systems often without sufficient scientific evidence of their efficacy and the lack of regulatory oversight. Both might be held especially problematic when this method is used for toys, games, and educational aids specifically targeted at children with ADHD. The authors conclude by pointing out the importance of further targeted research in this specific field and the need for responsible communication.

Fabrice Jotterand and James Giordano address potential challenges in using BCIs for the assessment and treatment of psychopathy. Given the lack of effective psychopharmacological treatments for psychopathy, the authors welcome the opportunity to investigate the possible utility of BCI technology in this field. A brief introduction to psychopathy and the limited diagnostic tools for these patients is followed by a discussion of potential ethical concerns with an emphasis on questions of personal identity.

Conclusion and Future Directions

Brain-computer interfacing presents exiting research directions at the overlap of neuroscience, micro-electronics, nanotechnology and the clinic. While BCIs promise unprecedented insight in and a better understanding of the brain and it functions they also raise important ethical questions. The ethical questions discussed in the contributions to this section present a comprehensive overview of the ethics of BCI technology to accompany its development. Neuroscience together with ethical considerations will enable to realize the full potential of this technology inside and outside the clinic to improve quality of life and provide innovative directions in arts, gaming and probably other areas of daily living.

Cross-References

▶ Attention Deficit Hyperactivity Disorder: Improving Performance Through Brain–Computer Interface
▶ Brain–Machine Interfaces for Communication in Complete Paralysis: Ethical Implications and Challenges
▶ Consciousness and Agency
▶ Detecting Levels of Consciousness
▶ Determinism and Its Relevance to the Free-Will Question
▶ Ethical Issues in Auditory Prostheses
▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
▶ Free Will and Experimental Philosophy: An Intervention
▶ No Excuses: Performance Mistakes in Morality
▶ Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges
▶ Sensory Enhancement

References

Baldwin, T., Fitzgerald, M., Kitzinger, J., Laurie, G., Price, J., Rose, N., Rose, S., Singh, I., Walsh, V., & Warwick, K. (2013). *Novel neurotechnologies: Intervening in the brain*. London: Nuffield Council on Bioethics.

Berger, T. W., Chapin, J. K., Gerhardt, G. A., McFarland, D. J., Principe, J. C., Soussou, W. V., Taylor, D. M., & Tresco, P. A. (2007). *WTEC panel report on international assessment of research and development in brain-computer interfaces*. Baltimore: WTEC.

Birbaumer, N., & Cohen, L. G. (2007). Brain-computer interfaces: communication and restoration of movement in paralysis. *J Physiol, 579*, 621–636.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kubler, A., Perelmouter, J., Taub, E., & Flor, H. (1999). A spelling device for the paralysed. *Nature, 398*, 297–298.

Buch, E., Weber, C., Cohen, L. G., Braun, C., Dimyan, M. A., Ard, T., Mellinger, J., Caria, A., Soekadar, S., Fourkas, A., & Birbaumer, N. (2008). Think to move: A neuromagnetic brain-computer interface (BCI) system for chronic stroke. *Stroke, 39*, 910–917.

Clausen, J. (2006). Ethische Aspekte von Gehirn-Computer-Schnittstellen in motorischen Neuroprothesen. *Int Rev Inf Ethics, 5*, 25–32.

Clausen, J. (2008). Moving minds: Ethical aspects of neural motor prostheses. *Biotechnology Journal, 3*, 1493–1501.

Clausen, J. (2009). Man, machine and in between. *Nature, 457*, 1080–1081.

Clausen, J. (2011). Conceptual and ethical issues with brain-hardware devices. *Current Opinion in Psychiatry, 24*, 495–501.

Clausen, J. (2013). Bonding brains to machines: Ethical implications of electroceuticals for the human brain. *Neuroethics, 6*, 429–434.

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., McMorland, A. J., Velliste, M., Boninger, M. L., & Schwartz, A. B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet, 381*, 557–564.

Galán, F., Nuttin, M., Lew, E., Ferrez, P. W., Vanacker, G., Philips, J., & Millan, J. R. (2008). A brain-actuated wheelchair: Asynchronous and non-invasive Brain-computer interfaces for continuous control of robots. *Clinical Neurophysiology, 119*, 2159–2169.

Haselager, P., Vlek, R., Hill, J., & Nijboer, F. (2009). A note on ethical aspects of BCI. *Neural Networks, 22*, 1352–1357.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., van der Smagt, P., & Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature, 485*, 372–375.

Johansson, V. (2008). Do brain machine interfaces on nano scale pose new ethical challenges? In J. S. Ach & C. Weidemann (Eds.), *Size matters: Ethical, legal and social aspects of nanobiotechnology and nano-medicine* (pp. 75–99). Münster: LIT.

Leeb, R., Friedman, D., Müller-Putz, G. R., Scherer, R., Slater, M., & Pfurtscheller, G. (2007). Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: A case study with a tetraplegic. *Computational Intelligence and Neuroscience, 2007*, 79642, 1–8 .

Liao, L. D., Chen, C. Y., Wang, I. J., Chen, S. F., Li, S. Y., Chen, B. W., Chang, J. Y., & Lin, C. T. (2012). Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors. *Journal of Neuroengineering and Rehabilitation, 9*, 5.

Lucivero, F., & Tamburrini, G. (2008). Ethical monitoring of brain-machine interfaces: A note on personal identity and autonomy. *Artif Intell Soc, 22*, 449–460.

Mak, J. N., & Wolpaw, J. R. (2009). Clinical applications of brain-computer interfaces: Current state and future prospects. *IEEE Reviews in Biomedical Engineering, 2*, 187–199.

Munssinger, J. I., Halder, S., Kleih, S. C., Furdea, A., Raco, V., Hosle, A., & Kubler, A. (2010). Brain painting: First evaluation of a new brain-computer interface application with ALS-patients and healthy volunteers. *Frontiers in Neuroscience, 4*, 182.

Tamburrini, G. (2009). Brain to computer communication: Ethical perspectives on interaction models. *Neuroethics, 2*, 137–149.

van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., Haselager, P., Vuurpijl, L., Gielen, S., & Desain, P. (2009). The brain-computer interface cycle. *Journal of Neural Engineering, 6*, 041001.

Vlek, R. J., Steines, D., Szibbo, D., Kubler, A., Schneider, M. J., Haselager, P., & Nijboer, F. (2012). Ethical issues in brain-computer interface research, development, and dissemination. *Journal of Neurologic Physical Therapy, 36*, 94–99.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113*, 767–791.

Zickler, C., Halder, S., Kleih, S. C., Herbert, C., & Kubler, A. (2013). Brain painting: Usability testing according to the user-centered design in end users with severe motor paralysis. *Artificial Intelligence in Medicine, 59*, 99–110.

Brain–Machine Interfaces for Communication in Complete Paralysis: Ethical Implications and Challenges

44

Surjo R. Soekadar and Niels Birbaumer

Contents

Abstract

Recent advances in computational capacities and sensor technologies allow online translation of electric, magnetic, or metabolic brain activity into control signals of computers or external devices. These brain-machine or brain-computer interfaces (BMI/BCI) allow individuals with complete paralysis to sustain communication, for example, enabling them to answer simple "yes-no" questions or to select letters in order to spell out whole words and phrases. The possibility to sustain communication through BMI/BCI systems in paralysis raises many critical neuroethical questions addressed in this chapter.

S.R. Soekadar (✉)
Department of Psychiatry and Psychotherapy, Applied Neurotechnology Lab/Institute of Medical Psychology and Behavioral Neurobiology, University Hospital Tübingen, Tübingen, Germany
e-mail: surjo.soekadar@uni-tuebingen.de; surjo@soekadar.com

N. Birbaumer
Institute of Medical Psychology and Behavioral Neurobiology, Tübingen, Germany

IRCCS, Ospedale San Camillo, Istituto di Ricovero e Cura a Carattere Scientifico, Lido di Venezia, Italy
e-mail: niels.birbaumer@uni-tuebingen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_41,

After an introduction and overview of the available BMI/BCI systems used for communication, ethical implications of this novel technology are drafted and discussed in direct reference to results and new insights from several clinical studies. These suggest, for example, that quality of life in locked-in syndrome (LIS), a state in which an individual is unable to move or speak, is not as limited and poor as generally believed, so that "patient wills" or "advanced directives" that were signed long before the locked-in state are questionable and might be used to avoid financial and social burdens by shortening what is anticipated to be long periods of care. After discussing limitations and challenges of the current BMI/BCI technology, the chapter closes with a future outlook and perspectives.

Introduction

Since its first conception by Jacques Vidal in the early 1970s (Vidal 1973), much work has advanced toward realization of direct brain communication via brain-computer or brain-machine interfaces (BCI/BMI), allowing communication without the involvement of the body. Following the ideas of Austrian novelist Ludwig Hohl (1904–1980) that "the human being lives according to its communication capacity" (Hohl 1984), the possibility of disembodied communication, which extends conventional means of how humans interact with their environment, might appeal to our fantasies of grandeur, but becomes existential once these conventional means have been lost, as "loosing the capacity for communication means loosing life" (Hohl 1984).

The development and applicability of BMI/BCI systems has been primarily driven by improvements in computational capacities and sensor technology, allowing for the instant translation of electric, magnetic, or metabolic brain activity into control signals or commands of computers and external devices (Birbaumer and Cohen 2007). The abbreviations BCI and BMI are synonymously used in the literature; however, authors dealing with neural interfaces controlling prosthetic devices or robots generally prefer the latter (Wolpaw and Wolpaw 2012). It is conceivable that the next few decades will see substantial shifts of current technical frontiers, while some concrete scientific problems are still unsolved. It is foreseeable, though, that the feasibility to communicate with severely paralyzed or behaviorally unresponsive individuals entails ethical implications, urging for a broader interdisciplinary discourse.

Such discourse, however, requires some basic understanding of the technical and scientific state-of-the-art, and must consider medical, legal, and often also cultural aspects of the societies in which this technology is used.

The first BCI systems providing real-time detection of brain events date back to the late 1970s (e.g., Vidal 1977) and take advantage of the early observation that electric brain potentials and oscillations first described by Hans Berger (1929) that show event- or task-related modulations, e.g., in preparation of a movement (Gastaut 1952), when solving a mental task (Klimesch 1996) or in preparation or reaction to an external stimulus (Chapman and Bragdon 1964). These modulations were topographically specific, so that the modality of brain function that was "engaged" during a specific

time interval could be identified. In 1999, Niels Birbaumer and his group at the University of Tübingen were the first to use such a system, which they called "thought translation device" (TTD) in two patients suffering from amyotrophic lateral sclerosis (ALS) (Birbaumer et al. 1999), a neurological disorder that leads to degradation of upper and lower motor neurons, often resulting in complete paralysis. These patients suffered from locked-in syndrome (LIS), a term introduced by Plum and Posner in 1966, referring to a neurological condition characterized by quadriplegia, lower cranial nerve paralysis, and the inability to talk, whereas consciousness and the ability to perform eye blinks remains intact (Plum and Posner 1966). By providing an alternative communication channel for this paralyzed patient, availability of BCI systems raised notable enthusiasm that even when the patient's ability to perform eye blinks or any other muscle twitches has vanished, a state termed complete locked-in syndrome (CLIS), such technology would still help sustain the patient's ability to communicate (Wolpaw et al. 2000). This perspective seemed particularly relevant as larger studies with individuals in LIS who were chronically respirated and artificially fed indicated that their quality of life was higher than expected (Lulé et al. 2013; Kübler et al. 2005a, 2006, Kübler and Birbaumer 2008) and their will to live largely uncompromised.

But where are we today? Despite innumerous experiments resulting in hundreds of scientific papers, no single scientifically documented patient with complete locked-in syndrome (CLIS) has ever been able to re-establish communication using a BMI/BCI. This suggests a more fundamental theoretical problem of learning and attention in advanced stages of ALS, as formulated in the extinction-of-thought hypothesis. At the same time, a larger multicenter clinical study using functional magnetic resonance imaging (fMRI), a technique to quantify task-related changes of blood-oxygenation level dependent (BOLD) signals, indicates that the rate of misdiagnosis in patients with disorders of consciousness ranges at approximately 30–40 % (Monti et al. 2010). Patients who are behaviorally unresponsive and become diagnosed as being in a vegetative or minimally conscious state show signs of wakefulness such as periodic eye opening and closing, but no signs of awareness. However, this recent study has demonstrated that they might in fact be well aware of their environment, and thus misdiagnosed. Recent work has shown that, for example, two patients with traumatic head injuries who were diagnosed as being in a minimally conscious state (MCS) and one patient diagnosed as in a vegetative state (VS) for 12 years were able to guide their attention to repeatedly communicate correct answers to binary "yes" or "no" questions by modulating their metabolic brain activity (Naci and Owen 2013).

These and other study results have brought to light a variety of scientific and ethical problems that were not expected before. For example: Are patients, relatives, and care-givers sufficiently informed about quality of life in complete paralysis and the possibilities of sustained communication via BMI/BCI technology when formulating advanced directives or patient wills? Which steps are important in the process of informed consent related to the application of BMI/BCI technology? Can a patient perform legal acts or ultimately decide on ending life support using a BMI/BCI system? How should we deal with advanced directives signed long before the present state of a disease?

After giving a general overview of currently established invasive and noninvasive BMI/BCI systems for communication and summarizing several clinical studies, this chapter strives to introduce some major ethical challenges, opening a discussion on possible perspectives on how these new technologies might fit into daily clinical practice.

Brain-Machine Interfaces for Communication

BMI/BCI systems rely on the detection, classification, and interpretation of brain activity, which can be recorded from sensors inside or outside the skull. Depending on the site of recording, BMI/BCI can be categorized as invasive or noninvasive systems (Wolpaw 2007). Other categorizations relate to the specific brain signal used for BMI/BCI control or the mode of operation. In terms of operation mode, active, passive, and reactive BMI/BCI applications can be distinguished. While active and reactive (also "evoked") BMI/BCI require the user's full attention to generate voluntary and directed commands, passive BMI/BCI relates to the concept of cognitive monitoring, introducing the assessment of the users' intentions, situational interpretations, and emotional states (Zander and Jatzev 2012).

In active BMI/BCI applications, two forms of control can be distinguished: synchronous and asynchronous control (Müller-Putz et al. 2006). In synchronous control, translation of brain activity follows a fixed sequence or cue, often indicated by external stimuli. Brain control is only possible during these fixed time intervals, and users are asked to rest in between the control trials. The main advantage of this design is that it allows for continuous calibration of the BCI/BMI system so that "active" and "rest" states can be reliably discriminated, even if, due to the non-stationarity of brain states, features of brain physiology vary during rest, e.g., due to fatigue. In asynchronous or un-cued control, the user has continuous control. Both operation modes have advantages and disadvantages depending on the purpose of the use, and it is conceivable that even a combination might be desirable in certain applications. For instance, using an asynchronous "brain switch" to start BMI/BCI control (based on a non-ambiguous brain signal that does not require continuous calibration or comparison with an unstable "rest" state), followed by a synchronous mode for communication (Pan et al. 2013).

The following paragraphs will introduce the best established and tested BMI/BCIs, with particular emphasis on those that were already successfully used for communication in patients with paralysis.

Noninvasive BMI/BCI

Since the discovery that electric or magnetic brain oscillations contain information about cognitive states (Ray and Cole 1985), which can be functionally specific (Chatrian et al. 1959; Walsh 1953), the idea to use such signals for direct brain control of assistive machines has become a major driving force for the development of BMI/BCI technology (Wolpaw et al. 2002).

The following noninvasively recorded neurophysiological signals were successfully used to control assistive devices in individuals with paralysis: (1) slow cortical potentials (SCPs) (Birbaumer et al. 1990, 1992), (2) sensorimotor rhythms (SMR) and their harmonics (Kübler et al. 2005b; Soekadar et al. 2011), and (3) event-related potentials (ERPs), e.g., P300 (Sellers and Donchin 2006; Nijboer et al. 2008) or steady-state visual evoked potential (SSVEP) (Müller-Putz et al. 2005).

Use of SCP in BMI/BCI applications goes back to Niels Birbaumer and his coworkers' work in the late 1970s showing that operant control of SCPs in healthy volunteers is possible (Lutzenberger et al. 1982). Later, such control enabled a paralyzed ALS patient to select single letters on a computer screen spelling to write entire text paragraphs (Birbaumer et al. 1999, 2003; Perelmouter and Birbaumer 2000; Wolpaw et al. 2002). Up to now, over 40 individuals with ALS at various stages of their disease have been trained to use such SCP-BCIs, with 7 of them in LIS. However, attempts to train a person after transitioning into CLIS, once any voluntary muscle twitches vanished, were unsuccessful (Hinterberger et al. 2005; Wilhelm et al. 2006; DeMassari et al. 2013a). The finding that any instructions to have specific thoughts or imaginations did not result in any measurable physiological (brain) response later gave reason to the extinction of thought hypothesis (Kübler and Birbaumer 2008; Birbaumer et al. 2014).

In contrast to SCPs, sensorimotor rhythms (SMRs) are recorded over the sensorimotor cortex usually at a frequency of 8–15 Hz. Depending on the context, SMR is also termed μ-rhythm or rolandic alpha, and disappears (*desynchronizes*) during planned, actual, or imagined movements (Gastaut 1952; Howe and Sterman 1972). SMR modulation during a given time interval can be quantified as event-related desynchronization (ERD) and synchronization (ERS) (Pfurtscheller and Aranibar 1979), a measure that later became extensively used for synchronous (externally cued) BMI/BCI control. While SMR *desynchronization* (the relative decrease of synchronously active neuronal cell assemblies) seems to reflect widespread information processing within the sensory-motor system (Leocani et al. 2001), SMR *synchronization* seems to reflect synchronous idling of sensory-motor networks (Pfurtscheller et al. 1996). Use of SMR-modulation for BMI/BCI control was extensively investigated by the Pfurtscheller group in Graz (Pfurtscheller and Neuper 2006), the Wolpaw group in Albany (Wolpaw and McFarland 2004; Wolpaw 2007), and the Birbaumer group in Tübingen (Mellinger et al. 2007; Soekadar et al. 2011).

Another well-established and tested brain signal used for BMI/BCI control is the event-related potential (ERP), e.g., the steady-state visual evoked potential (SSVEP) (Müller-Putz et al. 2005) or P300 as introduced by Donchin (Farwell and Donchin 1988). ERP-BCIs quantify the response of the brain to an external stimulus ("externally evoked response"). While SCP and SMR control often requires multiple training sessions for reliable control, ERP-BCIs basically do not require any training. In the classical P300-ERP-BCI paradigm, a letter matrix is presented on a visual screen. The user is instructed to select one of the characters by focusing on it, while the letters start flashing in a random order. Every time the selected letter flashes, a P300 response (a positive potential shift recordable

approximately 300 ms after the stimulus) can be detected, allowing to infer which of the letters was selected by the user. Besides visual stimuli, also other sensory qualities such as tactile (van der Waal et al. 2012) or auditory stimuli (Furdea et al. 2009; Schreuder et al. 2010) were successfully implemented in P300-ERP-BCI. SSVEP-BCI evaluates the continuous visual cortical response evoked by repetitive stimuli of a specific frequency. As perception of flickering visual stimuli results in SSVEP of the same frequency and its harmonics, evaluation of SSVEP provides information about the user's visual attention, e.g., whether the user is attending the flickering stimuli or not. In such paradigm, detection of an SSVEP at a specific frequency becomes linked to a certain computer command. For instance, an SSVEP at 20 Hz indicates opening of a robotic hand and an SSVEP at 7 Hz closing. By giving the user the possibility to fixate one of two LED displays flickering at 20 and 7 Hz, respectively, evaluation of the SSVEP allows for one to infer whether the user intends to open or to close the robotic arm (Sakurada et al. 2013).

Information transfer rates (ITR) can reach over 100 bits/min using SSVEP-BMI/BCI, or 20–30 bits/min in ERP-BMI/BCI, compared to 1–5 bits/min in SCP- and SMR-BMI/BCI systems. ERP-BMI/BCI systems depend on an intact sensory system. Most neurological disorders that lead to LIS or CLIS, however, are associated with degradation of the sensory capacities (Weis et al. 2011; Lulé et al. 2010). For instance, most ALS patients in more advanced stages of the disease suffer from compromised eyes which are constantly dry over time, impeding the applicability of assistive systems based on eye gaze control (Lorenceau 2012) or pupil responses (Stoll et al. 2013). For these individuals, pre-assessment of the intactness of afferent pathways is essential before any attempt to use ERP-based BMI/BCI systems. The auditory modality seems the most promising modality after onset of complete paralysis, although auditory BMIs are underrepresented in the BMI/BCI research field as their performance is below visual-BCI-tasks (Nijboer et al. 2008).

Most recently, a blood oxygenation level dependent (BOLD) signal-based real-time fMRI (rtfMRI) system has been introduced for BMI/BCI control (Weiskopf et al. 2003; Yoo et al. 2004; deCharms et al. 2004; Caria et al. 2012). As intracortical activity is highly correlated with local blood flow change and the BOLD signal (Logothetis et al. 2001), such signal changes associated with differential cortical activations, e.g., during motor imagery, can be used for affirmative or negative brain responses following specific questions or letter spelling (Sorger et al. 2012).

One of the most appealing features of brain metabolic responses, such as BOLD, for brain communication in LIS is the superior learning curve when controlling these responses compared to neuroelectric or neuromagnetic activity. This might relate to the ability for more specific targeting of functional areas using metabolic BMIs compared to noninvasively recorded electric or magnetic activity that reflects activity of large and widespread cell-assemblies. In other words, identification of task-related BOLD responses seems more accessible due to their relatively high functional specificity compared to identification of event- or task-related activity from ongoing and chaotic electromagnetic activity. Another explanation is that the mechanisms underlying self-regulation of vascular changes within brain tissue

differ from those involved in modulation of electro-chemical activity and are more accessible to volitional control.

While SCP regulation and selection of letters on a computer screen based on SCP changes in LIS (Birbaumer et al. 1999) and patients with intractable epilepsy required 30–100 sessions of intensive skill training, control of BOLD response was acquired within a few trials. For instance, rtfMRI of the cortex (deCharms et al. 2004, 2005), Sorger et al. (2012)) showed spelling of all letters of the alphabet with differential imagery using rtfMRI-based control, an achievement currently unthinkable with neuroelectric responses.

In addition to the rtfMRI-BCI approach, near-infrared spectroscopy (NIRS) is also a noninvasive technique based on measuring metabolic changes of the brain. Using multiple pairs or channels of light sources and light detectors operating at two or more discrete wavelengths at near-infrared range (700–1,000 nm), cerebral oxygenation and blood flow of particular regions of the cortical surface can be determined. The degree of increase in regional cerebral blood flow (rCBF) exceeds that of the increase in regional cerebral oxygen metabolic rate (rCMRO2), resulting in a decrease in deoxygenated hemoglobin in venous blood during higher oxygen demand. Therefore, an increase in total hemoglobin and oxygenated hemoglobin with a decrease in deoxygenated hemoglobin can be measured in activated brain areas. The recent development of portable systems makes NIRS a promising tool in noninvasive BMI/BCI research (Sitaram et al. 2007). Such a system might be of particular relevance for bedside testing and applying BMI/BCI systems for communication (Birbaumer et al. 2014).

Invasive BMI/BCI

The current research on invasive BMI/BCI mainly steers toward restoration of motor function in patients with quadriplegia to partly overcome their motor impairments (Collinger et al. 2013; Velliste et al. 2008; Hochberg et al. 2006), and less toward restoration of communication in locked-in patients. While invasive BMI/BCI has received the majority of the scientific and media attention, its significance for dealing with clinical problems related to LIS or complete paralysis is minor so far. It is unclear whether invasive BMI/BCI technology provides a reasonable solution with which these patients can communicate.

It is unquestioned that population codes of neural cell assemblies, represented in the firing patterns of its single cells, constitute the substrate for behaviorally relevant information. Whether it is the spiking of cells or the synaptically generated local field potentials (LFP) that encode this information is debatable (see Abeles 1991). Nevertheless, if all population codes of behaviorally relevant cell assemblies are recorded with an implanted electrode array, such invasive BMI/BCI would offer the largest degrees of freedom. Translation of neuronal spiking with linear algorithms recorded with multi-electrode implants in the motor cortex (Nicolelis 2012) or parietal cortex (Velliste et al. 2008) has allowed realistic reconstruction of functional hand and arm movements in 3D-space. So far, the BMI/BCI performance

accuracy of these systems cannot be reached with noninvasive electro- or magnetoencephalography (EEG/MEG) based BMI/BCI systems. However, decoding of two- to three-graded movement dimensions such as reaching and grasping or right-left movements using noninvasive MEG recordings has been demonstrated (Waldert et al. 2008). Hochberg et al. reported on human individuals who received a 100-microelectrode implant placed in their motor cortex and who were able to move and grasp with a robotic arm (Hochberg et al. 2006, 2012). These patients, though, were neither locked-in nor completely paralyzed, and not in need of a BMI/BCI device as this type of control of a robotic arm or computer cursor could be realized by using intact eye or head muscles.

Controlled studies are required that investigate the necessity of an invasive procedure for any therapeutic or assistive application of a BMI/BCI. Risk-benefit ratio is yet to be systematically evaluated. Versatile high-dimensional control of external devices acquired within a few hours or days represents a major advantage of invasive BMI/BCI systems, while their relevance for application in restoration of communication in LIS or complete paralysis remains uncertain.

A series of BMI/BCI studies on human patients with subdurally or epidurally implanted macroelectrodes recording ECoG in presurgically implanted patients with epilepsy (Leuthardt et al. 2004; Hinterberger et al. 2008) showed high classification rates of 70–90 % accuracy in selecting letters from different speller systems using brain oscillations derived from motor-related areas with frequencies up to the high gamma range without extensive training times. Patients used different types of imagery to select or ignore a particular letter or object from a computer menu. These studies are of experimental interest only, because none of these patients needed a BMI/BCI for communication or movement. However, BMI/BCI control is usually better with ECoG than EEG for obvious biophysical reasons. It has to be shown, however, whether invasive BMI/BCI will offer a more reliable and reasonable option for restoration of communication than any of the available noninvasive options.

Current Challenges

Besides missing evidence that any BMI/BCI can be reliably controlled after onset or transition into CLIS, a substantial challenge is the high rate of false positive and false negative classifications during BMI/BCI control. False positive classifications lead to unintended commands that impede the possible use and applicability of BMI/BCI systems. Useful communication (based on simple “yes” or “no” answers) needs at least 70 % correct classification of brain signals over an extended period of sessions (Perelmouter and Birbaumer 2000). The limited reliability of any of the available BCI/BMI systems has important normative implications, e.g., in the context of informed consent or any BCI/BMI-based communication with important consequences for the patient, e.g., medical or legal decisions. Another problem is that the vigilance and attention span of paralyzed individuals is often highly fluctuating.

After failure to communicate with individuals in CLIS using EEG-based BMI/BCI and other measures (e.g., pH-recordings (Wilhelm et al. 2006)), it was tested whether improving the quality and degrees of freedom in the brain signals (e.g., by using single cell firing or LFP after implantation of an electrocorticogram (ECoG)) would allow for better communication. ECoG electrodes were implanted in two individuals diagnosed with an advanced stage of ALS. However, none of these two patients achieved reliable brain communication rates with ECoG-BMIs (Birbaumer 2006).

This essentially negative result of invasive brain recording in CLIS suggests a more fundamental theoretical problem of learning and attention in brain communication with BMI.

Extinction of Thought

A meta-analysis of all reported ALS patients by Kübler and Birbaumer (2008) and a longitudinal study by Silvoni et al. (2013) indicate that no patient in CLIS has ever achieved sufficient communication rates and BMI/BCI performance meeting a 70 % correct threshold for sufficient on-line classification of "yes" or "no" responses. This threshold was tested by using questions with definite answers (e.g., "your name is Anton").

Although individuals in CLIS would benefit the most from direct brain communication, only a few investigations have aimed at finding reasons for the failure to achieve reliable BCI control in this population. It has been hypothesized that losing reinforcement contingencies between behavior and its feedback leads to extinction of goal-directed thinking and voluntary intentions. In complete paralysis, thoughts or motor intentions are not followed by their anticipated consequences. This would result, over a period of time, in extinction of goal-directed thinking itself (Birbaumer 2006; Kübler and Birbaumer 2008). Any attempts to use operant conditioning of the human brain after onset of complete paralysis (e.g., in complete locked-in syndrome (CLIS)) would fail even if afferent input and cognitive processing (attention, memory, imagery) remained intact.

If the voluntary response is only cognitive, such as in covert goal-orientated imagery in the locked-in state, the feedback or reward does not follow a reliable environmental or internal change and consequently extinguishes. Psychophysical studies (Haggard et al. 2002) demonstrate that if the behavioral response is elicited independently of a conscious decision and intention, the conscious awareness of the contingency and the conscious experience of the decision (the 'will') vanish. However, it is conceivable that CLIS patients use internally imagined response-effect contingencies that allow for the maintenance of goal directed thinking.

The hypothesis of extinction of reward-based learning in paralyzed organisms is supported by findings of experiments on curarized rats (Dworkin and Miller 1986). These studies have demonstrated the inability of operant (voluntary) learning to control visceral functions under complete paralysis. As the life-sustaining bodily functions of curarized rats were kept constant in these experiments, the homeostatic effect of the reward (rewarding brain stimulation or avoidance from shock) on body functions and, as a consequence, on learning was absent.

However, Dworkin showed that despite the absence of instrumental learning, classical conditioning of curarized rats was possible (Dworkin 1993). After pairing

tones with aversive stimuli, they learned to control autonomic functions such as blood pressure, vasoconstriction, and heart rate in response to the tones (Dworkin and Dworkin 1995). Following this conceptual framework, Razran, a student of Pavlov, demonstrated classical *semantic* conditioning (Razran 1961). In such a paradigm, words or sentences belonging to different semantic categories are presented as a conditioned stimulus (CS). One specific semantic category (i.e., the thought "no") is followed by an unconditioned stimulus (US), usually a non-painful electric shock. Razran showed semantic conditioning of saliva production to words with positive valence and found evidence for transfer to synonyms, but not to homophonic words. In the same way, he demonstrated generalization to sentences with identical contextual statements or even emotional valence, independent of the constituent words of the sentences (Razran 1949, 1961). Previous studies of cortical correlates of differential semantic classical conditioning have shown an increased amplitude of the evoked brain responses (event-related potentials, ERPs) following the onset of the conditioned stimulus (CS; pseudowords or syllables) predictive of an aversive event (Montoya et al. 1996; Heim and Keil 2006). Following these findings, Birbaumer and his team developed a classical semantic conditioning BMI/BCI allowing online classification of different components of EEG oscillations, ERPs and NIRS signals associated to "yes" or "no"-responses in such a classical semantic conditioning paradigm (Furdea et al. 2012; Ruf et al. 2013; DeMassari et al. 2013b).

Obviously, classical semantic conditioning of autonomic responses or brain responses does not depend on goal-directed motor systems and involves only minimal attentional resources and effort. Thus, classical semantic conditioning may circumvent extinction of volition, goal-directed thinking, effortful selective attention, cognitive control, imagery, and working memory functions during operant learning and, thus, represent a principle to maintain communication in complete paralysis. By implementing a classical semantic conditioning paradigm into a metabolic NIRS-based BMI, a first successful case of communication with a CLIS patient was reported recently (Birbaumer et al. 2014).

A particular challenge is the validation of answers to open questions, as validation often depends on family members and long-term nursing personnel who likely introduce biases toward their own expectations. Frequent repetition of questions that are semantically differently phrased (i.e. "you are in constant pain," instead of "you are NOT in constant pain") might help to control for this problem.

In summary, first studies suggest that classical semantic conditioning might be a possibility to overcome the "silence" after onset of CLIS, particularly when using metabolic brain signals due to the superior learning to control these brain signals. However, replication of these findings in a larger group of CLIS patients is necessary.

Vigilance and Fragmentation of Sleep

Finally, LIS and CLIS caused by destruction or degeneration of the motor system and the connected frontal areas in ALS, or due to brain stem stroke or other brain damage, is usually accompanied by changes in circadian rhythm, sleep pattern, and arousal with alterations of attentional performance (Soekadar et al. 2013).

These changes may at least partially be responsible for failing to establish BMI/BCI communication. While in LIS (locked-in state) with intact eye-control and other states of incomplete paralysis, circadian rhythm is largely preserved, sleep polysomnographic recordings over many nights in one CLIS patient with implanted ECoG-electrodes showed increasing fragmentation of slow wave sleep (SWS), even during day time hours (Soekadar et al. 2013). BMI sessions usually are scheduled during daytime hours, without polysomnographic recordings and without on-line detection of sleep. Fading vigilance during learning of BMI-based communication methods may be an inevitable consequence, because sleep cannot be detected behaviorally by any means in a motionless, eyes-closed, and artificially respirated CLIS patient, except through analysis of EEG signals.

The relationship between vigilance and (possibly any) BMI/BCI performance was also substantiated in several post-hoc analyses of available data recorded in LIS patients, indicating an inverse correlation (ranging from 0.4 to 0.7) between the reduction of P300 amplitudes (assessed during an "auditory odd-ball task" requiring patients to count specific rare auditory stimuli) and BMI/BCI performance based on semantic conditioning across multiple sessions (DeMassari et al. 2013b).

Ethical Dimensions of Brain-Computer Interfaces for Communication

The ethical dimensions of BMI/BCI technology for communication in paralysis can be divided into the following areas: (1) ethical aspects related to advanced directives and the process of informed consent toward the use of assistive technologies and life-sustaining measures, (2) ethical aspects directly related to the use of such technology, and (3) ethical aspects related to the availability and access to such technology.

The issue of informed consent, defined as the process in which participants or legal representatives become fully informed about the risks, potential benefits, and costs of a medical treatment, procedure, or clinical trial, is regulated in most countries (e.g., the Directive 2001/20/EC in the European Union or the Code of Federal Regulations, CFR, Title 21, Part 50.20 in the United States of America). Application of BMI/BCI systems in this regard does not differ from use of any other assistive or supportive technology in patient populations. However, proper discourse over the cost-benefit ratios requires involvement of multiple parties with complementary expertise, particularly to avoid having end-of-life plans that are changed based on false hopes and misinformation. Multiple strategies and guidelines were formulated to assist clinicians and researchers in this process (Haselager et al. 2009).

Current legal systems in most industrialized countries provide the possibility to declare a 'patient will' or 'advanced directive' in case of medical emergencies. The patients' "advanced directive" or "patient will" can include the wish that "artificial" life-preserving medical interventions such as artificial respiration and feeding and similar intensive care measures would be prohibitive and should be avoided. In most cases, however, these patients' wills are signed long before such medical

emergency occurs. The underlying assumption of the signing persons (and these documents) is very often that the circumstances requiring life-preserving measures are associated with a poor quality of life.

In contrast to a large body of studies, the widespread assumption prevails that quality of life in advanced stages of ALS, chronic stroke, and other chronic and untreatable neurological disorders associated with paralysis and dependence of daily assistance or life sustaining measures is poor. The opposite is the case: several longitudinal studies indicate that quality of life in such populations is average and sometimes even above average despite paralysis, artificial respiration, and feeding (Lulé et al. 2013; 2009; Kübler et al. 2005a, 2006, Kübler und Birbaumer 2008). It was found that paralyzed and respirated ALS patients may judge their quality of life as satisfactory and good, while others consider the motionless individual to have a very poor quality of life (Kübler et al. 2005a; Lulé et al. 2009, 2012; Matuz et al. 2010; Pantke and Birbaumer 2012). Despite these results, about 95 % of all individuals diagnosed with ALS "choose" to die before tracheostomy and artificial respiration (Borasio 1996; Neudert et al. 2001). These studies suggest that patients' will-declarations or advanced directives may be invalid at the decisive time points. They may serve as an argument for the medical professions, insurance companies, or family members to shorten anticipated long periods of care for these patients and avoid the financial and social burdens. Patients themselves may follow the social pressure and accept the poor quality-of-life prognosis and ask for hastened death.

The ethical aspects directly related to the use of BMI/BCI technology concern the expected consequences on the quality of life of the users (and their relatives/caregivers), possible risks and side-effects, lack of reliability that might lead to erroneous interpretation of the user's intention (with normative implications for issues such as informed consent or life-changing decisions), and also include other issues like human dignity, mental and bodily integrity, personhood, and communication to the media (Nijboer et al. 2011; Clausen 2011). These aspects are partly subject of dynamic technological advancements, and in part identical to issues raised in the context of the applications of other neurotechnological tools, such as deep brain stimulation (DBS) or cochlea implants.

There is one dimension, however, that becomes increasingly important given that BMI/BCI can be used to diagnose contextual awareness and enable completely paralyzed patients to communicate. In acute locked-in syndrome (LIS), eye-coded communication and evaluation of cognitive and emotional functioning is very limited, because vigilance is fluctuating and eye movements may be inconsistent, very small, and easily exhausted, leading to the provisional diagnosis of CLIS. It has been shown that more than half of the time it is the family and not the physician who first realized that the patient was aware. Recent studies reported that the diagnosis of LIS on average takes over 2.5 months. In some cases, it took 4–6 years before aware and sensitive patients, locked in an immobile body, were recognized as being conscious (Laureys et al. 2005). Once a LIS patient becomes medically stable, and given appropriate medical care, life expectancy increases to several decades. Even if the chances of good motor recovery are very limited, existing eye-controlled computer-based communication technology currently

allows the patient to control his environment, use a word processor coupled to a speech synthesizer, and access the internet (Caligari et al. 2013).

The availability of such assistive technology, including BMI/BCI systems, may force us to disregard death requests formulated in advanced directives or patient wills that were signed long before the present state of a disease. It might urge for the formulation of guidelines and more differentiated courses of action in the case of onset of complete locked-in syndrome (CLIS) and the case that recovery of communication is unsuccessful over a specified period of time.

Knowledge about the possibility of contextual awareness in the unresponsive and recovery of communication despite complete paralysis (though unresolved) requires a broader discourse about the regular application of EEG monitoring and use of BMI/BCI systems in such populations, even if first attempts to restore communication fail.

This raises the issue of availability, access, and proficiency to use such technology within the medical systems. Given the existential dimension and preventable suffering of individuals who are cognitively intact and aware but have no means to express such awareness, a problem of increasing relevance due to demographic developments, e.g., rising incidence of stroke and improved survival rates after traumatic brain injury (TBI), society needs to decide on how it wants to face this reality to avoid that "loosing the capacity for communication means loosing life" (Hohl 1984).

Based on the available studies presented in this chapter dealing with the application of BMI/BCI technology for communication, the following points seem to be of particular relevance:

1. The diagnosis of a progressive disease that might lead to CLIS requires a thorough discussion with the patient and his/her relatives and caregivers about the quality of life for individuals in a state of complete paralysis.
2. BMI/BCI might become a technical option to sustain or restore communication for individuals in CLIS, but more studies are needed that involve such patients. Up until now, no scientifically tested BMI/BCI systems provided reliable communication in CLIS.
3. Given the high rate of misdiagnosed individuals with contextual awareness despite the diagnosis of a vegetative state (VS) (Naci and Owen 2013; Cruse et al. 2011), repeated testing of unresponsive individuals seems advisable. This also applies to patients in CLIS, as their vigilance might be highly fluctuating (Soekadar et al. 2013) or their awareness might be compromised by episodes of minimally conscious state (MCS).

Future Outlook and Conclusions

Since the presence of a patient's awareness significantly impacts multiple ethical dimensions including social, medical, and legal issues related to surrogate decisions on end-of-life care and others, there are an increasing number of people advocating for mandatory additional testing of covert awareness in those diagnosed as

behaviorally unresponsive (Kotchoubey et al. 2013; Cruse et al. 2011). Besides the possibility for repeated attempts to detect task-related changes associated with motor imagery, the ability to employ long-term EEG monitoring would allow quantification of vigilance possibly affected by increased sleep fragmentation and disturbed circadian rhythm (Soekadar et al. 2013).

Innovative paradigms that are particularly promising for communication with complete locked-in syndrome (CLIS) patients based on classical semantic conditioning (DeMassari et al. 2013b) might be particularly useful to detect covert awareness in the behaviorally unresponsive and to restore communication. Application of different approaches ranging from eye-gaze or pupil response-based assistive systems to adequate BMI paradigms based on the combination of different brain signals (e.g., using the combination of NIRS and EEG) might improve the reliability of BMI/BCI systems for communication. In this context, the assessment of the individual's "decision making capacity" might become a relevant aspect, e.g., when involving a patient into meaningful clinical decision-making or legal processes (Peterson et al. 2013).

From a clinical and an ethical perspective, a proper and repeated evaluation and diagnosis of patients' state of consciousness whenever in question, especially in those who are behaviorally unresponsive, seems advisable or even mandatory.

Recent advances in technology have allowed for communication with severely paralyzed individuals (e.g., those suffering from locked-in syndrome or paresis after stroke or brain injury). Multiple clinical studies indicate that these patients may, in fact, have a higher quality of life than otherwise assumed by outsiders, or than they expected themselves before entering a specific state of paralysis. In addition, recent studies have shown that some of those who have previously been perceived as behaviorally unresponsive and in vegetative state were misdiagnosed and consciously aware. While these findings are a consequence of the availability of novel technologies, such as BMI/BCI systems, they raise several ethical questions, for example, the role of advanced directives, which are written long before an individual is in a certain state. In addition, improving reliability and applicability of these tools in completely paralyzed or behaviorally unresponsive patients represents a major scientific and medical challenge. There are many indicators, though, that this challenge can be mastered, given the necessary political will. Particularly, access to these technologies and how they should be used are questions that will require intensified interdisciplinary discourse and interactions between patients, caregivers, physicians, engineers, and scientists, as our ability to sustain, preserve, and enhance human life continually improves.

Acknowledgments We thank Sook-Lei Liew and Birgit Teufel for their help in preparing this manuscript. This work was supported by the German Federal Ministry of Education and Research (BMBF, 16SV5840 and 01GQ0831), the European Commission under the project WAY (#288551), the Deutsche Forschungsgemeinschaft (DFG SO932-2, Reinhart Koselleck Project), the Volkswagen Stiftung, the BrainProducts GmbH, and the Baden-Württemberg Stiftung gGmbH.

Cross-References

▶ Consciousness and Agency
▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
▶ Human Brain Research and Ethics

References

Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

Berger, H. (1929). Ueber das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten, 87*, 527–570.

Birbaumer, N. (2006). Breaking the silence: Brain-computer-interfaces (BCI) for communication and motor control. *Psychophysiology, 43*, 517–532.

Birbaumer, N., & Cohen, L. (2007). Brain-computer-interfaces (BCI): Communication and restoration of movement in paralysis. *The Journal of Physiology, 579*, 621–636.

Birbaumer, N., Elbert, T., Canavan, A., & Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behavior. *Physiological Reviews, 70*, 1–41.

Birbaumer, N., Gallegos-Ayala, G., Wildgruber, M., Silvoni, S., & Soekadar, S. R. (2014). Direct brain control and communication in paralysis. *Brain Topography, 27*, 4–11.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., & Flor, H. (1999). A spelling device for the paralysed. *Nature, 398*, 297–298.

Birbaumer, N., Hinterberger, T., Kübler, A., & Neumann, N. (2003). The thought-translation device (TTD): Neurobehavioral mechanisms and clinical outcome. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*, 120–123.

Birbaumer, N., Roberts, L. E., Lutzenberger, W., Rockstroh, B., & Elbert, T. (1992). Area-specific self-regulation of slow cortical potentials on the sagittal midline and its effects on behavior. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 84*, 353–361.

Borasio, G. D. (1996). Discontinuing ventilation of patients with amyotrophic lateral sclerosis. Medical, legal and ethical aspects. *Medizinische Klinik, 91*, 51–52.

Caligari, M., Godi, M., Guglielmetti, S., Franchignoni, F., & Nardone, A. (2013). Eye tracking communication devices in amyotrophic lateral sclerosis: Impact on disability and quality of life. *Amyotroph Lateral Scler Frontotemporal Degener, 14*, 546–552 [Epub ahead of print].

Caria, A., Sitaram, R., & Birbaumer, N. (2012). Real-time fMRI: A tool for local brain regulation. *The Neuroscientist, 18*, 487–501.

Chapman, R. M., & Bragdon, H. R. (1964). Evoked responses to numerical and non-numerical visual stimuli while problem solving. *Nature, 203*, 1155–1157.

Chatrian, G. E., Petersen, M. C., & Lazarte, J. A. (1959). The blocking of the rolandic wicket rhythm and some central changes related to movement. *Electroencephalography and Clinical Neurophysiology, 11*, 497–510.

Clausen, J. (2011). Conceptual and ethical issues with brain-hardware interfaces. *Current Opinion in Psychiatry, 24*, 495–501.

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., McMorland, A. J., Velliste, M., Boninger, M. L., & Schwartz, A. B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet, 381*, 557–564.

Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T. A., Fernández-Espejo, D., Pickard, J. D., Laureys, S., & Owen, A. M. (2011). Bedside detection of awareness in the vegetative state: A cohort study. *Lancet, 378*, 2088–2094.

deCharms, R. C., Christoff, K., Glover, G. H., Pauly, J. M., Whitfield, S., & Gabrieli, J. D. (2004). Learned regulation of spatially localized brain activation using real-time fMRI. *NeuroImage, 21*, 436–443.

deCharms, R. C., Maeda, F., Glover, G. H., Ludlow, D., Pauly, J. M., Soneji, D., Gabrieli, J. D. E., & Mackey, S. C. (2005). Control over brain activation and pain learned by using real-time functional MRI. *Neuriscience, PNAS, 102*, 18626–18631.

DeMassari, D., Matuz, T., Furdea, A., Ruf, C. A., Halder, S., & Birbaumer, N. (2013b). Brain-computer interface and classical conditioning of communication in paralysis. *Biological Psychology, 92*, 267–274.

DeMassari, D., Ruf, C. A., Furdea, A., Matuz, T., van der Heiden, L., Halder, S., Silvoni, S., & Birbaumer, N. (2013a). Brain communication in the locked-in state. *Brain, 136*, 1989–2000.

Dworkin, B. R. (1993). *Learning and physiological regulation*. Chicago: The University of Chicago Press.

Dworkin, B. R., & Dworkin, S. (1995). Learning of physiological responses: II. Classical conditioning of the baroreflex. *Behavioral Neuroscience, 109*, 1119–1136.

Dworkin, B. R., & Miller, N. E. (1986). Failure to replicate visceral learning in the acute curarized rat preparation. *Behavioral Neuroscience, 100*, 299–314.

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 70*, 510–523.

Furdea, A., Halder, S., Krusienski, D. J., Bross, D., Nijboer, F., Birbaumer, N., & Kübler, A. (2009). An auditory oddball (P300) spelling system for brain-computer interfaces. *Psychophysiology, 46*, 617–625.

Furdea, A., Ruf, C., Halder, S., DeMassari, D., Bogdan, M., Rosenstiel, W., Matuz, T., & Birbaumer, N. (2012). A new (semantic) reflexive brain-computer interface: In search for a suitable classifier. *Journal of Neuroscience Methods, 203*, 233–240.

Gastaut, H. (1952). Electrocorticographic study of the reactivity of rolandic rhythm. *Review Neurologique (Paris), 87*, 176–182.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*, 382–385.

Haselager, P., Vlek, R., Hill, J., & Nijboer, F. (2009). A note on ethical aspects of BCI. *Neural Networks, 22*, 1352–1357.

Heim, S., & Keil, A. (2006). Effects of classical conditioning on identification and cortical processing of speech syllables. *Experimental Brain Research, 175*, 411–424.

Hinterberger, T., Birbaumer, N., & Flor, H. (2005). Assessment of cognitive function and communication ability in a completely locked-in patient. *Neurology, 64*, 1307–1308.

Hinterberger, T., Widmann, G., Lal, T. N., Hill, J., Tangermann, M., Rosenstiel, W., Schölkopf, B., Elger, C., & Birbaumer, N. (2008). Voluntary brain regulation and communication with electrocorticogram signals. *Epilepsy & Behavior, 13*, 300–306.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., et al. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature, 485*, 372–375.

Hochberg, L. R., Serruya, M. D., Friehs, G. M., et al. (2006). Neuronal ensemble control of prostheticdevices by a humanwithtetraplegio. *Nature, 442*, 164–171.

Hohl, L. (1984). Notizen Suhrkamp, Frankfurt am Main, ISBN 3-518-37500-8, 64.

Howe, R. C., & Sterman, M. B. (1972). Cortical-subcortical EEG correlates of suppressed motor behavior during sleep and waking in the cat. J. *Electroencephalography and Clinical Neurophysiology, 32*, 681–695.

Klimesch, W. (1996). Memory processes, brain oscillations and EEG synchronization. *International Journal of Psychophysiology, 24*, 61–100.

Kotchoubey, B., Veser, S., Real, R., Herbert, C., Lang, S., & Kübler, A. (2013). Towards a more precise neurophysiological assessment of cognitive functions in patients with disorders of consciousness. *Restorative Neurology and Neuroscience, 31*, 473–485.

Kübler, A., & Birbaumer, N. (2008). Brain-computer interfaces and communication in paralysis: Extinction of goal directed thinking in completely paralysed patients? *Clinical Neurophysiology, 119*, 2658–2666.

Kübler, A., Nijboer, F., Mellinger, J., Vaughan, T. M., Pawelzik, H., Schalk, G., McFarland, D. J., Birbaumer, N., & Wolpaw, J. R. (2005b). Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface. *Neurology, 64*, 1775–1777.

Kübler, A., Weber, C., & Birbaumer, N. (2006). Locked-in – Freigegeben für den Tod? Wenn nur Denken und Fühlen bleiben – Neuroethik des Eingeschlossenseins. *Zeitschrift für medizinische Ethik, 52*, 57–70.

Kübler, A., Winter, S., Ludolph, A. C., Hautzinger, M., & Birbaumer, N. (2005a). Severity of depressive symptoms and quality of life in patients with amyotrophic lateral sclerosis. *Neurorehabilitation and Neural Repair, 19*, 182–193.

Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., Berré, J., Faymonville, M. E., Pantke, K. H., Damas, F., Lamy, M., Moonen, G., & Goldman, S. (2005). The locked-in syndrome: What is it like to be conscious but paralyzed and voiceless? *Progress in Brain Research, 150*, 495–511.

Leocani, L., Toro, C., Zhuang, P., Gerloff, C., & Hallet, M. (2001). Event-related desynchronization in reaction time paradigms: A comparison with event-related potentials and corticospinal excitability. *Clinical Neurophysiology, 112*, 923–930.

Leuthardt, E. C., Schalk, B., Wolpaw, J. R., Ojemann, J. G., & Moran, D. W. (2004). A brain-computer interface using electrocorticographic signals in humans. *Journal of Neural Engineering, 1*, 63–71.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature, 412*, 150–157.

Lorenceau, J. (2012). Cursive writing with smooth pursuit eye movements. *Current Biology, 22*, 1506–1509.

Lulé, D., Ehlich, B., Lang, D., Sorg, S., Heimrath, J., Kübler, A., Birbaumer, N., & Ludolph, A. C. (2013). Quality of life in fatal disease: The flawed judgement of the social environment. *Journal of Neurology, 260*, 2836–2843.

Lulé, D., Diekmann, V., Müller, H. P., Kassubek, J., Ludolph, A. C., & Birbaumer, N. (2010). Neuroimaging of multimodal sensory stimulation in amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry, 81*, 899–906.

Lulé, D., Pauli, S., Altintas, E., Singer, U., Merk, T., Uttner, I., Birbaumer, N., & Ludolph, A. (2012). Emotional adjustment in amyotrophic lateral sclerosis (ALS). *Journal of Neurology, 259*, 334–341.

Lulé, D., Zickler, C., Bruno, M. A., Demertzi, A., Pellas, F., Laureys, S., & Kübler, A. (2009). Life can be worth living in locked-in syndrome. *Progress in Brain Research, 177*, 339–351.

Lutzenberger, W., Elbert, T., Rockstroh, B., & Birbaumer, N. (1982). Biofeedback produced slow brain potentials and task performance. *Biological Psychology, 14*, 99–111.

Matuz, T., Birbaumer, N., Hautzinger, M., & Kubler, A. (2010). Coping with amyotrophic lateral sclerosis: An integrative view. *Journal of Neurology, Neurosurgery, and Psychiatry, 81*, 893–898.

Mellinger, J., Schalk, G., Braun, C., Preissl, H., Rosenstiel, W., Birbaumer, N., & Kübler, A. (2007). An MEG-based brain-computer interface (BCI). *NeuroImage, 36*, 581–593.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful modulation of brain activity in disorders of consciousness. *The New England Journal of Medicine, 362*, 579–589.

Montoya, P., Larbig, W., Pulvermüller, F., Flor, H., & Birbaumer, N. (1996). Cortical correlates of semantic classical conditioning. *Psychophysiology, 33*, 644–649.

Müller-Putz, G. R., Scherer, R., Brauneis, C., & Pfurtscheller, G. (2005). Steady-state visual evoked potential (SSVEP)-based communication: Impact of harmonic frequency components. *Journal of Neural Engineering, 2*, 123–130.

Müller-Putz, G. R., Scherer, R., Pfurtscheller, G., & Rupp, R. (2006). Brain-computer interfaces for control of neuroprostheses: From synchronous to asynchronous mode of operation. *Biomedizinische Technik (Berlin), 51*, 57–63.

Naci, L., & Owen, A. M. (2013). Making every word count for nonresponsive patients. *JAMA Neurology*. doi:10.1001/jamaneurol.2013.3686.

Neudert, C., Oliver, D., Wasner, M., & Borasio, G. D. (2001). The course of the terminal phase in patients with amyotrophic lateral sclerosis. *Journal of Neurology, 248*, 612–616.

Nicolelis, M. A. (2012). Mind in motion. *Scientific American, 307*, 58–63.

Nijboer, F., Clausen, J., Allison, B. Z., & Haselager, P. (2011). Researchers' opinions about ethically sound dissemination of BCI research to the public media. *International Journal of Bioelectromagnetism, 13*, 108–109.

Nijboer, F., Sellers, E. W., Mellinger, J., Jordan, M. A., Matuz, T., Furdea, A., Halder, S., Mochty, U., Krusienski, D. J., Vaughan, T. M., Wolpaw, J. R., Birbaumer, N., & Kübler, A. (2008). A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. *Clinical Neurophysiology, 119*, 1909–1916.

Pan, J., Li, Y., Zhang, R., Gu, Z., & Li, F. (2013). Discrimination between control and idle states in asynchronous SSVEP-based brain switches: A pseudo-key-based approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 21*, 435–443.

Pantke, K.-H., & Birbaumer, N. (2012). Die Lebensqualität Schwerstbetroffener nach einem Schlaganfall mit Locked-in Syndrom. *Logos Interdisziplinär, 20*, 296–300.

Perelmouter, J., & Birbaumer, N. (2000). A binary spelling interface with random errors. *IEEE Transactions on Rehabilitation Engineering, 8*, 227–232.

Peterson, A., Naci, L., Weijer, C., Cruse, D., Fernández-Espejo, D., Graham, M., & Owen, A. M. (2013). Assessing decision-making capacity in the behaviorally nonresponsive patient with residual covert awareness. *AJOB Neuroscience, 4*, 3–14.

Pfurtscheller, G., & Aranibar, A. (1979). Evaluation of event-related desynchronization (ERD) preceding and following self-paced movement. *Electroencephalography Clinical Neurophysiology, 46*, 138–146.

Pfurtscheller, G., & Neuper, C. (2006). Future prospects of ERD/ERS in the context of brain-computer interface (BCI) developments. *Progress in Brain Research, 159*, 433–437.

Pfurtscheller, G., Stancák, A., Jr., & Neuper, C. (1996). Event-related synchronization (ERS) in the alpha band–an electrophysiological correlate of cortical idling: A review. *International Journal of Psychophysiology, 24*, 39–46.

Plum, F., & Posner, J. B. (1966). *The diagnosis of stupor and coma*. Philadelphia: F. A. Davis.

Ray, W. J., & Cole, H. W. (1985). EEG activity during cognitive processing: Influence of attentional factors. *International Journal of Psychophysiology, 3*, 43–48.

Razran, G. (1949). Semantic and phonetographic generalizations of salivary conditioning to verbal stimuli. *Journal of Experimental Psychology, 39*, 642–652.

Razran, G. (1961). The observable unconscious and the inferable conscious in current soviet psychophysiology: Interoceptive conditioning, semantic conditioning, and the orienting reflex. *Psychological Review, 68*, 1–147.

Ruf, C., DeMassari, D., Wagner-Podmaniczky, F., Matuz, T., & Birbaumer, N. (2013). Semantic conditioning of salivary pH for communication. Special Issues, *Artificial Intelligence in Medicine, 59*, 91–98.

Sakurada, T., Kawase, T., Takano, K., Komatsu, T., & Kansaku, K. (2013). A BMI-based occupational therapy assist suit: Asynchronous control by SSVEP. *Frontiers in Neuroscience, 7*, 172.

Schreuder, M., Blankertz, B., & Tangermann, M. (2010). A new auditory multi-class brain-computer interface paradigm: Spatial hearing as an informative cue. *PLoS ONE, 5*(4), e9813. doi:10.1371/journal.pone.0009813.

Sellers, E. W., & Donchin, E. (2006). A P300-based brain-computer interface: Initial tests by ALS patients. *Clinical Neurophysiology, 117*, 538–548.

Silvoni, S., Cavinato, M., Volpato, C., Ruf, C. A., Birbaumer, N., & Piccione, F. (2013). Amyotrophic lateral sclerosis progression and stability of brain-computer interface communication. *Amyotrophic Lateral Sclerosis Frontotemporal Degeneration, 14*, 390–396.

Sitaram, R., Guan, C., Zhang, H., Thulasidas, M., Hoshi, Y., Ishikawa, A., Shimizu, K., & Birbaumer, N. (2007). Temporal classification of multi-channel near infrared spectroscopy signals of motor imagery for developing a brain-computer interface. *NeuroImage, 34*, 1416–1427.

Soekadar, S. R., Born, J., Birbaumer, N., Bensch, M., Halder, S., Murguialday, A. R., Gharabaghi, A., Nijboer, F., Schölkopf, B., & Martens, S. (2013). Fragmentation of slow wave sleep after onset of complete locked-in state. *Journal of Clinical Sleep Medicine, 9*, 951–953.

Soekadar, S. R., Witkowski, M., Mellinger, J., Ramos, A., Birbaumer, N., & Cohen, L. G. (2011). ERD-based online brain-machine interfaces (BMI) in the context of neurorehabilitation: optimizing BMI learning and performance. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 19*, 542–549.

Sorger, B., Reithler, J., Dahmen, B., & Goebel, R. (2012). A real-time fMRI-based spelling device immediately enabling robust motor-independent communication. *Current Biology, 22*, 1333–1338.

Stoll, J., Chatelle, C., Carter, O., Koch, C., Laureys, S., & Einhäuser, W. (2013). Pupil responses allow communication in locked-in syndrome patients. *Current Biology, 23*, 647–648.

van der Waal, M., Severens, M., Geuze, J., & Desain, P. (2012). Introducing the tactile speller: An ERP-based brain-computer interface for communication. *Journal of Neural Engineering, 9*, 045002.

Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S., & Schwartz, A. B. (2008). Cortical control of a prosthetic arm for self-feeding. *Nature, 453*, 1098–1101.

Vidal, J. J. (1977). Real-time detection of brain events in EEG. *Proceedings of the IEEE, 65*, 633–641.

Vidal, J. J. (1973). Toward direct brain-computer communication. *Annual Review of Biophysics and Bioengineering, 2*, 157–180.

Waldert, S., Preissl, H., Demandt, E., Braun, C., Birbaumer, N., Aertsen, A., & Mehring, C. (2008). Hand movement direction decoded from MEG and EEG. *Journal of Neuroscience, 28*, 1000–1008.

Walsh, E. G. (1953). Visual attention and the alpha-rhythm. *Journal of Physiology, 120*, 155–159.

Weis, J., Katona, I., Müller-Newen, G., Sommer, C., Necula, G., Hendrich, C., Ludolph, A. C., & Sperfeld, A. D. (2011). Small-fiber neuropathy in patients with ALS. *Neurology, 76*, 2024–2029.

Weiskopf, N., Veit, R., Erb, M., Mathiak, K., Grodd, W., Goebel, R., & Birbaumer, N. (2003). Physiological self-regulation of regional brain activity using real-time functional magnetic resonance imaging (fMRI): Methodology and exemplary data. *NeuroImage, 19*, 577–586.

Wilhelm, B., Jordan, M., & Birbaumer, N. (2006). Communication in locked-in syndrome: Effects of imagery on salivary pH. *Neurology, 67*, 534–535.

Wolpaw, J. R. (2007). Brain-computer interfaces as new brain output pathways. *Journal of Physiology, 579*, 613–619.

Wolpaw, J. R., & McFarland, D. J. (2004). Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 17849–17854.

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., & Vaughan, T. M. (2000). Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering, 8*, 164–173.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113*, 767–791.

Wolpaw, J., & Wolpaw, E. (2012). Brain-computer interfaces: Something new under the sun. In Wolpaw J. R. & Wolpaw E. W. (Eds), *Brain-computer interfaces: Principles and practice* 1–5. Oxford University Press.

Yoo, S. S., Fairneny, T., Chen, N. K., Choo, S. E., Panych, L. P., Park, H., Lee, S. Y., & Jolesz, F. A. (2004). Braincomputer interface using fMRI: Spatial navigation by thoughts. *NeuroReport, 15*, 1591–1595.

Zander, T. O., & Jatzev, S. (2012). Context-aware brain-computer interfaces: Exploring the information space of user, technical system and environment. *Journal of Neural Engineering, 9*, 016003.

Ethical Issues in Brain–Computer Interface Research and Systems for Motor Control

45

Donatella Mattia and Guglielmo Tamburrini

Contents

Abstract

BCI (brain-computer interface) systems for motor control process patterns of brain activity and translate them into signals which enable one to replace, restore, or improve impaired motor capabilities. BCI operation involves sustained interactions of the central nervous system with computational and robotic systems that are themselves capable of adaptive behaviors. This fact molds distinctively BCI ethical issues. Indeed, experimental BCI therapies for motor rehabilitation target directly brain areas and raise special informed consent and medical beneficence issues in view of the fact that the functional

D. Mattia (✉)
Clinical Neurophysiology, Neuroelectrical Imaging and BCI Laboratory, Fondazione Santa Lucia IRCCS, Rome, Italy
e-mail: d.mattia@hsantalucia.it

G. Tamburrini
DIETI – Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Università di Napoli Federico II, Naples, Italy
e-mail: tamburrini@unina.it

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_42,

implications of brain plasticity are not fully understood and are difficult to predict. Similarly, epistemic limitations concerning the future behaviors of adaptive BCI systems shape ethical reflection on retrospective distribution of responsibilities and liabilities for damages caused by BCI-actuated devices.

Novel ethical issues arise in connection with more distant prospects for BCI enhancement of unimpaired motor capabilities. Ethical policy formation about BCI-enabled enhancements appears to be premature in view of technological lack of imminence. Nevertheless, watchful monitoring of BCI research is presently called for, in order to anticipate prospective ethical tensions between the claims of personal freedom to enhancement and the claims deriving from social justice, fairness, and mental and physical integrity considerations.

On the whole, BCI systems afford unique potential solutions for protecting the autonomy, the action, and even the thinking capabilities of people affected by severe motor impairments. However, trust building between BCI researchers and various groups of stakeholders requires the development of communication strategies which enable one to appreciate the rapid advancements in BCI research without underestimating at the same time the formidable challenges one has to meet before various forms of BCI-enabled communication and motor control become more widely available.

Introduction

A brain-computer interface (BCI) processes brain activity online and identifies patterns of this activity that one uses for a variety of communication and motor control purposes. More recently, BCI systems have been used for brain-monitoring purposes too.

Motor control has been a major focus of BCI research from its very beginning. There, selected patterns of brain activity are processed and translated into signals which enable one to replace, restore, or improve impaired motor capabilities (Millán et al. 2011; Wolpaw and Walpow 2012, see Chaps. 1, 11, 22). For example, BCI-actuated prostheses, robotic manipulators, and wheelchairs may *replace* the functionalities of impaired or lost limbs. By delivering electrical impulses to the muscles of a paralyzed hand, BCI systems enable one to *restore* hand control when normal neuromuscular pathways are disrupted as a consequence of spinal cord injury. And BCI systems have been more recently used to monitor brain activity patterns in the course of poststroke rehabilitation therapies, with the aim of *improving* brain motor control and movement execution.

In addition to these various assistive and rehabilitative applications, one may envision prospective uses of BCI systems in the service of healthy people. BCI prototypes for the general population include systems for attention, emotional state, and mental workload monitoring, photograph or video archive sorting, and game playing. Current limitations of BCI communication channels in the way of

bit-rate transfer suggest that only special categories of healthy users may benefit in the near future from suitable BCI systems for motor control. It has been suggested that astronauts might be one of these categories, insofar as mental teleoperation enables one to govern external semiautomatic manipulators in microgravity working conditions (Millán et al. 2006).

A variety of ethical issues arise in connection with experimental research on BCI systems for motor control and their intended uses in both medical and nonmedical contexts. These issues are not unique to BCI systems: they are closely related to ethical themes emerging in other areas of neuroethics (Clausen 2009), and they are subsumed under standard categories of applied ethics, such as informed consent and respect for persons, medical beneficence, autonomy and liberty, responsibility and liability, distributive justice, and responsible communication of research results (Clausen 2008; Tamburrini 2009). Nevertheless, various distinctive traits of BCI ethics emerge from combinations of the following circumstances:

(i) *Brain-computer mutual adaptation*. BCI operation involves the interaction of two *adaptive* controlling units. These are the central nervous system (CNS) and the computer program which identifies features of brain activity and translates them into control signals.

(ii) *Machine intelligence*. BCI-controlled devices are endowed with varying degrees of intelligence and autonomous action capabilities. This design choice is usually motivated by compelling technological reasons.

(iii) *Potential users*. The groups of potential users who are more likely to take advantage in the near future from BCI systems are formed by persons affected by severe motor impairments: in particular, by those who are left with very limited or no residual capabilities for voluntary movement.

The ensuing reflections on (i)–(iii) prepare the ground for an examination of ethical issues that are shaped by various combinations of these items.

Brain-Computer Mutual Adaptation

Interaction between a BCI system and its user starts from the performance of some *mental task* on the part of the user. During task performance, the user's brain activity is *recorded* and *processed* online in order to *recognize* the presence of features that one *translates* into control signals for some external device. Finally, the user obtains a (perceptual or linguistic) *feedback* about the outcomes of this repeatable interaction cycle.

This closed-loop functional scheme can be multiply realized in ways that are conditional on available kinds of BCI hardware and software, on the invasive or noninvasive character of brain signal recording devices, on the use of synchronous or asynchronous interaction protocols, and on the variety of brain signals that are acquired and processed. The latter include sensorimotor rhythms or P300 event-related potentials (ERPs) among the electrical signals and BOLD (blood oxygen level dependent) responses among the metabolic signals. Additional taxonomies arise on the basis of a variety of informative distinctions, which comprise those

between dependent and independent, active and passive, and hybrid and non-hybrid BCI systems (Wolpaw et al. 2002; Wolpaw and Walpow 2012).

Mutual adaptations between CNS and BCI arise from information exchanges flowing in accordance with the basic functional organization of brain-machine interactions outlined above. These mutual adaptations occur in both preparatory stages and normal operating conditions.

To begin with, one should be careful to note that a central component of a BCI system is a computer program which recognizes some given number of brain activity patterns. For example, in order to set up a BCI system for controlling a robotic wheelchair (Millán et al. 2004), one has to develop a computer program which is capable of recognizing as many brain activity correlates of different mental tasks as one needs to issue a repertoire of basic motor commands, such as “turn left,” “turn right,” “go ahead,” and “stop.” These computational components of BCI systems are usually developed by means of *machine-learning* techniques and protocols (Müller et al. 2008), which crucially involve a *training* phase carried out on a sample of brain activity correlates of mental task performance. Since brain activity patterns vary across different subjects during the performance of the same mental task, this training phase is crucially needed to tune the machine up to its individual user, thereby achieving satisfactory classifications of his or her brain activity patterns.

The need for machine-to-human adaptation by learning may arise during normal operating conditions too, as soon as one observes significantly declining classification performances on the part of the machine. Deteriorating classification performances depend on various factors affecting brain correlates of mental task performance: fatigue, increased familiarity with mental tasks, changes induced by biological clocks, and variable attention levels are notable cases in point.

Initial machine-to-human adaptations are followed by human-to-machine adaptations throughout BCI normal operation. As a matter of fact, an operant conditioning process usually intervenes to change brain activity patterns whenever the user obtains negative feedback information, that is, whenever the user receives notification of discrepancies between expected and actual outcomes of her interaction with the machine. As a result of operant conditioning, brain activity patterns change in ways that have been found to facilitate ensuing machine classifications.

The basic functional organization of BCI interactions is modifiable by taking advantage of additional machine-to-human adaptation processes. The detection of so-called error potentials is a significant case in point. Error potentials are components of ERPs associated to the brain processing of errors, which arise a few hundred milliseconds after presentation of perceptual stimuli that the CNS will recognize as errors (Hoffmann and Falkenstein 2011). A BCI which learns to identify error potentials may undertake a self-correcting action before its human user is able to notify the error to the machine or to issue some correction command by means of the usual BCI communication channel (Buttfield et al. 2006).

Machine Intelligence

In noninvasive BCI systems, the need for some degree of autonomous intelligent action on the part of BCI-actuated devices, such as robotic wheelchairs or manipulators, chiefly arises from current limitations in bit-rate transfer (Wolpaw et al. 2002). In view of these limitations, users are confined to issue commands or convey chunks of information whose coding requires sequences of few bits only. For example, users may instruct a robotic wheelchair to move left or right, but they can hardly plan and control its left or right steering course in a more detailed fashion. Accordingly, one usually delegates lower-level control of left or right commands to the robotic wheelchair. Similarly, BCI-actuated robotic systems are entrusted with the job of meeting important deadlines that otherwise would be missed on account of the BCI bit-rate bottleneck. A pertinent example is the timely avoidance of unexpected obstacles on the trajectory of a robotic wheelchair. More in general, the use of BCI-actuated robots involves a division of labor between man and machine in the way of intelligent control: the user selects higher-level commands, and the machine, which is thereby granted some degree of autonomy, plans and carries out lower-level subtasks (Millán et al. 2004; Galán et al. 2008).

Classes of Potential Users

In view of the bit-rate bottleneck, current BCI systems – and especially the more widely used noninvasive systems – compare unfavorably with many alternative assistive technologies that are available to people who are able to make a variety of voluntary movements. For this reason, BCIs afford viable and often unique solutions for restricted groups of people only, that is, for those who are left with no or very limited movement capabilities. Indeed, early prototypes of BCI systems were conceived and developed for the benefit of people affected by the *classical* locked-in syndrome (LIS), which results in the complete paralysis of nearly all voluntary muscles except for eye moving and blinking (Farwell and Donchin 1988). One must be careful to emphasize, however, that effective use of BCI systems has not been demonstrated yet in the condition of *complete* LIS, that is, in the case of people who cannot perform any voluntary movement at all (Hochberg and Anderson 2012).

Consider now the impact on familiar ethical issues which derive from mutual man–machine adaptations in the BCI context, from the intelligence embedded into BCI-actuated devices, and from the special needs of users who are more likely to take advantage from BCI-enabled replacement of motor functions.

Informed Consent and Respect for Persons

Along with healthy participants, research trials on BCI-enabled replacement of motor functions may enrol people who have lost most of their abilities to

communicate and act. Clearly, these research trials are unlikely to bring personal advantages to disabled participants. Nevertheless, disabled participants and their families may come to view a BCI research trial as a unique and last resort to overcome excruciating communication and action barriers. These unrealistic expectations might be so compelling in the dramatic human condition of severely paralyzed persons that they prevent a proper appreciation of the facts one must know before participating in BCI experimentation (Haselager et al. 2009; Clausen 2011; Vlek et al. 2012). Accordingly, specific information aimed at anticipating and mitigating similar psychological reactions must be included in setting up informed consent questionnaires and protocols for BCI experimentation. In particular, one must provide clear and detailed information about the phenomenon of BCI illiteracy, that is, the incapability to operate a BCI, which is estimated to affect 15–30 % of potential users, and for which effective remedies are still to be found (Vidaurre and Blankertz 2010); about psychological risks of depression which may derive from retracting BCI use at the end of time-limited studies (Schneider et al. 2012); and about discomfort and complications in the care of disabled participants which may derive from prolonged operation and maintenance of BCI systems (Wolpaw and Walpow 2012, see Chap. 20).

It was pointed out above that the interaction of two adaptive controllers is crucially involved in BCI training and operation. As one shall see later on, distinctive risks arise in connection with unsatisfactory outcomes of machine-to-human adaptations. But ethically more significant concerns arise in connection with potentially deleterious changes in the CNS resulting from human-to-machine adaptations. Consider, from this perspective, BCI-enabled replacement of user motor functions. There, psychological rewards and punishments deriving from the observation of BCI interaction outcomes enable the CNS to learn and modify its activity so as to facilitate BCI actuation of user intents. Learning occurring in the BCI context is a source of brain plasticity (Wolpaw and Walpow 2012, see Chap. 13). Accordingly, as the overall functional implications of brain plasticity are not fully understood and difficult to predict, one cannot rule out generic risks of detrimental effects on states of mind and behaviors of BCI intensive users (Schneider et al. 2012). Ethically motivated safeguards that are similar to those adopted in connection with psychoactive medications (Merkel et al. 2007; Clausen 2009) or deep brain stimulation (DBS) devices (Clausen 2010) must be put in place to deal with similar risks. Neurologists, psychologists, and other medical personnel involved in BCI research and assistance to BCI users carry the moral duty of constant monitoring for timely detection of adverse consequences and risk assessment update. And every BCI candidate user must be properly informed of potentially detrimental effects of BCI-induced brain plasticity (Dobkin 2007).

Some BCI-based rehabilitation therapies for improving motor functions *target directly* brain areas, with the principal aim of modifying their structure and functions (Shih et al. 2012). This intended use of BCI systems differs from the use that one generally makes of these systems in order to replace lost motor functions. Indeed, BCI systems for controlling robotic wheelchairs do not aim directly at

changing brain structure: intervening modifications of this sort, if any, are a side effect of intensive BCI operation.

Various rehabilitation strategies have been made available to foster BCI-induced changes in the CNS. One strategy for poststroke motor rehabilitation, for example, involves a BCI system monitoring damaged motor areas of the brain which normally control movements that are impaired after brain injury. The BCI system provides appropriate feedback according to whether activation patterns in the targeted brain areas come closer to normal or not (Grosse-Wentrup et al. 2011; Pichiorri et al. 2011; Daly and Sitaram 2012; Mattia et al. 2012; Várkuti et al. 2013). One may use *hybrid* BCI systems for similar purposes (Bermudez et al. 2013): BMCI systems (with the letter "M" in the acronym standing for *muscle*) combine the analysis of brain signals and electromyographic signals to stimulate increasingly correct activity patterns in both brain and muscles.

Ethical concerns about BCI systems directly fostering brain plasticity arise, at a general level, from awareness of limited etiological understanding of deleterious effects, if any, of these interventions. This epistemic predicament calls for sustained monitoring of BCI-based therapies directly targeting the brain. However, no guidelines have been developed yet for risk management and sharing of data about these experimental therapies. Protection of the mental and physical integrity of patients provides a major ethical motivation for setting up research and task forces which aim at filling this gap. Candidate participants in these experimental therapies must be specifically informed of the fact that modifications in brain areas are a direct objective of these interventions rather than side effects resulting from intensive use of BCI systems.

More straightforward ethical recommendations concern informed consent about the use of *invasive* versus *noninvasive* BCI systems. Operation of invasive BCI systems requires the implant of electrodes in the cortex (Velliste et al. 2008; Hochberg et al. 2012; Collinger et al. 2013) or the deployment of apparatuses for electrocorticography (ECoG) or intracranial electroencephalography (iEEG) on the exposed brain surface (Moran 2010). Compared to noninvasive BCIs, invasive ones achieve better spatial signal resolution and signal-to-noise ratio leading to improved control of peripheral devices. Among the noninvasive systems, those relying on the EEG as a recording method afford better performances in terms of signal temporal resolution, cost, and practicality of use. Potential users of BCI systems must be informed of comparative advantages and disadvantages of invasive and noninvasive systems, including relevant facts about characteristic risks of invasive systems in connection with implant stability, reversibility, and infection. Interestingly, empirical data from interviews administered to people affected by amyotrophic lateral sclerosis (ALS) suggest a definite preference for noninvasive systems, notwithstanding the functional advantages of invasive BCIs which derive from better spatial resolution and signal-to-noise ratio, and the corresponding disadvantages of noninvasive systems in the way of slower operation and more error-prone control (Birbaumer 2006).

Beneficence, Justice, and Autonomy

International declarations on human rights furnish a very general ethical framework for addressing a variety of specific beneficence issues in bioethics. These documents notably include the United Nations Universal Declaration of Human Rights (United Nations 1948) and the more recent Charter of Fundamental Rights of the European Union (European Union 2000). A selective examination of these documents – let alone an examination of international treaties and conventions that are more specifically concerned with health-care rights – sheds some light on the operational content that BCI research and systems may bring to human rights promotion and protection. It is worth considering, from this vantage point, the potential contribution of BCI research and systems to the promotion and protection of human dignity.

Article 1 of the UN Declaration asserts that "All human beings are born free and equal in dignity and rights." And Article 1 of the EU Charter asserts that "Human dignity is inviolable. It must be respected and protected." Arguably, BCI systems for replacing impaired motor functions of severely paralyzed persons afford practical implementations of these principles via the conceptual intermediary of the notion of human agency. A human agent is an entity capable of performing a wide repertoire of actions guided by desires, intentions, and beliefs about the world. The vulnerability of human agency as such is dramatically underscored by the condition of people affected by LIS: almost completely lost action capabilities come there with the persisting mental capability to conceive intentions to act. The protection of human agency, which is compromised by LIS, plays a central role in major philosophical views of human dignity. According to Kantian views of dignity, for example, the inherent worth of a human being is ultimately grounded in the capability to act as *homo noumenon*, that is, in the capability to endorse rationally the moral maxims of practical reason and to conform one's own actions to those maxims (Rothaar 2010). And according to neo-Aristotelian views of dignity, the more direct way to realize a dignified human life is to develop and exercise human capabilities for practical deliberation, control over one's own environment, and engagement into social activities (Nussbaum 2006, pp. 77–78 and p. 161). Clearly, both conceptions of human dignity afford moral motivations for specific deeds which aim at protecting the action capabilities of people affected by severe motor disabilities. And these positive actions can be supported by a wide variety of BCI technological platforms, such as those for environmental control, typing, computer access and internet surfing, computer games, virtual drawing, and painting (Millán et al. 2011).

Consider now Article 26 of the EU Charter of Fundamental Rights: "The Union recognizes and respects the right of persons with disabilities to benefit from measures designed to ensure their independence, social and occupational integration and participation in the life of the community." There are different approaches to filling the gap between this general statement and positive actions that are inspired to it. In addition to shared views and political agreement about societal duties toward groups of disabled people, a crucial factor concerns the identification

of the more effective ways to use available resources. Thus, for example, technological research aimed at promoting the rights of disabled persons is encouraged with the crucial proviso of giving priority to the development of technologies at affordable costs (United Nations 2006, see in particular article 4, paragraphs f and g). Accordingly, scientists in the BCI community ought to pay attention to distributive justice concerns in their research work and to look for cost-effective BCI systems and technologies.

BCI systems might even come to play a role in the protection of persons and personhood as such. It was pointed out above that a successful demonstration of BCI use in the complete LIS condition is still missing. This learning inability may depend on a generalized decline of attention, perception, and thinking capabilities which is hypothesized to take over after the onset of complete LIS. According to an alternative explanation, however, sustained preservation of purposive thinking requires psychological reinforcements that occur only if the actual consequences of intended actions are verified. In the complete LIS condition, this reinforcement is occasionally forthcoming through the intermediary of caretakers who happen to fulfil the patient's current desire. Without reinforcement, purposive thinking fades away, and with it goes the capability to learn and operate a BCI. If this alternative explanation turns out to be correct, then one may prevent the extinction of goal-oriented thinking by teaching one how to operate a BCI before the onset of complete LIS (Birbaumer 2006). Accordingly, teaching a person in a state of incomplete or classical LIS how to use a BCI would serve the morally praiseworthy purpose of protecting personhood, to the extent that goal-oriented thinking is a central feature of a person.

Responsibility and Liability

In BCI systems for motor control, selected patterns of brain activity are recognized and translated into signals which enable one to replace, restore, or improve impaired motor capabilities. Thus, the proper functioning of BCI systems crucially requires reliable classifications of brain activity patterns. It was pointed out above that machine-learning methods are usually deployed to develop, on the basis of a variety of training techniques and data, computer programs which classify brain activity patterns. Various estimates of the reliability of learned classification rules can be worked out on both theoretical and experimental grounds. It is well known, however, that each of these estimates involves distinctive background assumptions about the significance of training data or the stability of the stochastic phenomena one is dealing with. Unfortunately, these background assumptions are difficult to buttress when the classification of brain activity patterns is at stake: increased familiarity with mental tasks, in addition to mental fatigue, mental changes induced by biological clocks, variable attention levels, and BCI system setup (such as the variable positioning of the EEG cap across training or normal operating sessions), is known to affect the stability of recorded brain signals. Briefly, the reliability of learned classification rules depends on boundary conditions on brain processing

that are difficult to isolate and control, insofar as task execution history, current mental context, and technical setup procedures concur to jeopardize the stability of recorded brain signals (Santoro et al. 2008). Discrepancies between user intents and actual robotic navigation trajectories may occur in BCI contexts of use even when user intents are correctly identified. These additional disagreements arise in view of uncertainties and predictive limitations which generally concern the behavior of robotic systems, such as robotic sensitivity to small perturbations of initial conditions or sensor noise piling up in series of sensory readings (Nehmzow 2006).

There are ethically relevant issues arising from known epistemic limitations about the behaviors of brain-actuated robots. The retrospective distribution of responsibilities and liabilities for damages caused by BCI-actuated robots is a case in point, insofar as programmers, manufacturers, and users are not in the position to predict exactly and certify what these robots will actually do in their intended operational environments (Tamburrini 2009; Clausen 2011). For example, how are responsibilities and liabilities sensibly distributed if a BCI-actuated robotic wheelchair rolled down a staircase due to its incorrect interpretation of user intents or when a robotic arm responding to a glass-of-water request hits another person standing in the room?

Programmers, manufacturers, and users who failed to predict some damaging event resulting from misclassification of user's intents, perceptual failures, or environmental perturbations cannot be held morally blameworthy, provided that they properly attended, in their different capacities, to every reasonable BCI design, implementation, testing, and operational issue. Nevertheless, even in the absence of moral responsibilities deriving from negligence or malevolent intentions, damaging events give rise to liabilities and objective responsibilities which require proper distribution.

In developing liability policies for brain-actuated robots, one may rely on a large body of legal norms and casuistry concerning liabilities for the use of machinery in general. One should be careful to note, however, that BCI-actuated systems stand out from machines of many other sorts, insofar as the symbol-processing, learning, reasoning, and action planning capabilities that BCI systems share to some extent with human beings and other biological systems are special sources of behavioral unpredictability in principle or in practice.

This positive analogy between the unpredictability of BCI systems and of biological systems suggests that the inability of BCI users to predict exactly and control the behavior of brain-actuated robots is meaningfully related to the inability of dog owners to curb their pets in every possible circumstance, to the inability of employers to predict exactly and control the behavior of their employees, and even to the inability of parents to predict and control the behavior of their children. Parents are nevertheless held to be vicariously liable for many kinds of damaging events caused by their children, just as pet owners are liable for damages caused by their pets, and employers are liable for certain types of damages caused by their employees. Accordingly, users should be held liable for damaging events resulting from unpredictable behaviors of their BCI systems. However, the suggestion of extending to BCI-actuated robots policies regulating liability ascription for

damages caused by unpredictable behaviors of human beings and other biological systems is not immune from ethically motivated criticism. Indeed, a liability policy of this kind may introduce discriminations in the access to assistive technologies between those who can and those who cannot afford the cost of insurance and compensations for damages caused by BCI-actuated devices.

Alternatively, one might shift the burden of economic compensations entirely on BCI manufacturers. In view of their expected profits, producers of goods are often held liable for damaging events that are difficult to predict and control. This liability ascription policy is aptly summarized in the Roman juridical tradition by means of the formula *ubi commoda, ibi incommoda*. However, ethically motivated criticism can be levelled against this alternative suggestion too. Indeed, the risk of high compensation costs may discourage investments in R&D toward the development of marketable BCI systems, with the effect of diverting resources that are much needed to start a pioneering BCI industry. As a consequence, tensions are likely to arise between a liability policy charging compensation costs on BCI producers, the demands of beneficence in bioethics, and consequentialist evaluations of societal benefits that are expected to flow from BCI technological innovation.

These various observations suggest that a complex governance framework, allowing for the socialization of risks associated to this innovative technology, is needed to address properly BCI liability problems. Since BCI technological risk comes with beneficial opportunities for groups of disabled people, in addition to longer-term benefits in the way of technological innovation for society at large, one might appropriately distribute insurance and compensation costs on a variety of stakeholder groups and governmental agencies. Similarly, the problem of identifying criteria for acceptable risk should be addressed within this complex governance framework.

Liberty

Astronauts, it was noted above, are a professional group of healthy users who might take advantage of BCI-actuated robots in view of microgravity conditions hindering their movements. In a more distant future, BCI systems for motor control might prove useful to enhance or extend the motor capabilities of wider groups of healthy people. These envisaged developments raise novel ethical issues, notably regarding the personal freedom to enhance one's own mental and physical capabilities vis-á-vis ethical concerns about mental and physical integrity, justice, and fair social opportunities. These ethical issues are quite novel. They are not equally pressing, however, according to any ethical triage which takes technological imminence in due account (Farah 2002; Tamburrini 2009; Clausen 2011). For this reason, attending now to ethically motivated norms and recommendations about envisioned BCI-enabled motor enhancements appears to be a premature move. Watchful ethical monitoring is presently more appropriate, so as to anticipate and issue early warnings about novel ethical issues, if any, arising from BCI research in this area (Lucivero and Tamburrini 2008). In particular, ethical monitoring is

needed to detect possible tensions between justice and equality on the one hand and personal freedom to use one's own economic resources to get BCI motor enhancement and extension on the other hand; to anticipate emerging pressures to get BCI motor enhancements in view of new job opportunities; and to understand how BCI motor enhancements might come to affect one's own personality, social relationships, and relationships with the natural environment at large.

Responsible Communication of BCI Research

In a democratic society, citizens and special groups of stakeholders responsibly participating in public affairs have to make up their minds about the uses of novel technologies and their coherence with given moral values and aspirations. On the basis of reflective and deliberative processes, individuals may form moral judgments and contribute to develop societal orientations or, more stringently, moral and legal norms about novel technologies. These processes are realistically bound to depend on a division of epistemic labor between ordinary citizens and experts, insofar as one cannot expect that ordinary citizens collect from scratch the background scientific and technological information which is needed to analyze and weigh viable alternatives (Kitcher 2011). Moral responsibilities are distributed accordingly. Citizens carry the moral responsibility to reduce their ignorance about scientific and technological matters which may threaten goals and aspirations that they value most; experts carry the moral responsibility of supplying what is, to their best knowledge, correct and comprehensive specialized information. There are, however, many ways in which experts intentionally or inadvertently produce misinformation about scientific and technological matters (Proctor 2008). Identifying and offsetting the underlying mechanisms is an ethically significant undertaking.

In the BCI context, public statements of researchers about their research goals and activities may become a powerful source of misinformation, insofar as disproportionate technological and therapeutic expectations are generated and amplified at the hands of both media reporting and individual psychological responses (Clausen 2008; Haselager et al. 2009). One might be able to countervail some of these undesirable outcomes of scientific communication by paying due attention to the distinction between the *general framework* of a research program and the series of detailed models, technologies, and systems that a research program brings progressively into being during its life course (Laudan 1978; Godfrey-Smith 2003). The general framework of a research program is put together by identifying basic assumptions about entities and processes in some domain of inquiry, theoretical ideas, and techniques which are heuristically identified as appropriate means to investigate phenomena there, in addition to overall research goals one hopes to attain in the long run. Clearly, the overall goals and promises of a general framework play significant motivating roles in scientific work and may suggest fruitful lines of inquiry, even when these promises cannot be attained without making substantial and unforeseeable progress with respect to currently available models and techniques.

In making public statements about their work, scientists cannot count on the shared background of tacit knowledge which shapes communication styles within their own scientific communities. Accordingly, the ethically praiseworthy goal of furnishing correct and accessible information to the general public raises the formidable challenge of adapting professional communication styles of scientists to the needs of broader audiences of nonspecialists. In particular, scientists should mark clearly up long-term or visionary goals of a research program, setting them apart from tangible results that have been actually achieved. Ethical motivations for carefully drawing these distinctions have emerged in connection with several developments of BCI research. In communicating the research goal of making a variety of BCI technologies and systems available to people affected by LIS, one ought to emphasize problems of BCI system reliability, cognitive decline in LIS patients, the incidence of BCI illiteracy, and the generalizability to clinical contexts of results obtained in research trials involving healthy subjects only. Moreover, one must specify BCI costs and benefits with respect to alternative communication methods for persons who retain some communication capabilities, say, by moving their eyes or eyelids.

The need for responsible communication strategies emerges even more evidently in connection with the suggestion of using BCI for communicating with and promoting the autonomy of people affected by disorders of consciousness. Similar suggestions were advanced in the wake of pioneering studies on communication protocols with persons diagnosed to be in a vegetative state: a limited number of these persons were able to answer autobiographical questions on the basis of a communication protocol involving the performance of a mental task and the analysis of its brain correlates (Owen et al. 2006; Monti et al. 2010). Extensions of this communication protocol were envisaged for the purpose of allowing subjective symptom reporting, informed consent, and other exchanges of information aiming at medical beneficence and autonomy protection. One should be careful to note, however, that assessing whether persons affected by disorders of consciousness still possess the required varieties and degrees of consciousness is a scientifically formidable and empirically elusive problem. In these circumstances, one ought to adopt an extremely cautious attitude in communicating the results of these studies and their possible developments to psychologically vulnerable families of people affected by disorders of consciousness (Tamburrini and Mattia 2011). More in general, trust building between BCI researchers and various groups of interested parties requires the development of suitable communication strategies which enable one to appreciate the rapid progresses made by BCI research without underestimating the formidable scientific and medical challenges that have to be met in the future before various forms of BCI-enabled communication and motor control become more widely available.

Cross-References

▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes

References

Bermudez, I., Badia, S., Garcia, M. A., Samaha, H., & Verschure, P. (2013). Using a hybrid brain computer interface and virtual reality system to monitor and promote cortical reorganization through motor activity and motor imagery training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 21*, 174–181. doi:10.1109/TNSRE.2012.2229295.

Birbaumer, N. (2006). Brain-computer interface research: Coming of age. *Clinical Neurophysiology, 117*, 479–483.

Buttfield, A., Ferrez, P. W., & Millán, J. del R. (2006). Towards a robust BCI: Error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14*, 164–168.

Clausen, J. (2008). Moving minds: Ethical aspects of neural motor prostheses. *Biotechnology Journal, 3*, 1493–1501.

Clausen, J. (2009). Man, machine, and in between. *Nature, 457*, 1080–1081.

Clausen, J. (2010). Ethical brain stimulation – Neuroethics of deep brain stimulation in research and clinical practice. *The European Journal of Neuroscience, 32*, 1152–1162.

Clausen, J. (2011). Conceptual and ethical issues with brain-hardware interfaces. *Current Opinion in Psychiatry, 24*, 495–501.

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., McMorland, A. J. C., Velliste, M., Boninger, M. L., & Schwartz, A. B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet, 381*, 557–564.

Daly, J. J., & Sitaram, R. (2012). BCI therapeutic applications for improving brain function. In J. R. Walpow & E. W. Walpow (Eds.), *Brain-computer interfaces: principles and practice* (pp. 351–362). Oxford: Oxford University Press.

Dobkin, B. H. (2007). Brain-computer interface technology as a tool to augment plasticity and outcomes for neurological rehabilitation. *The Journal of Physiology, 579*, 637–642.

European Union. (2000). Charter of fundamental rights of the European Union. http://www.europarl.europa.eu/charter/pdf/text_en.pdf

Farah, M. (2002). Emerging ethical issues in neuroscience. *Nature Neuroscience, 5*, 1123–1129.

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 70*, 510–523.

Galán, F., Nuttin, M., Lew, E., Ferrez, P. W., Vanacker, G., Philips, J., & Millán, J. del R. (2008). A brain-actuated wheelchair: Asynchronous and non-invasive brain-computer interfaces for continuous control of robots. *Clinical Neurophysiology, 119*, 2159–2169.

Godfrey-Smith, P. (2003). *Theory and reality. An introduction to the philosophy of science*. Chicago: University of Chicago Press.

Grosse-Wentrup, M., Mattia, D., & Oweiss, K. (2011). Using brain-computer interfaces to induce neural plasticity and restore function. *Journal of Neural Engineering, 8*. doi:10.1088/1741-2560/8/2/025004.

Haselager, P., Vlek, R., Hill, J., & Nijboer, F. (2009). A note on ethical aspects of BCI. *Neural Networks, 22*, 1352–1357.

Hochberg, L. R., & Anderson, K. D. (2012). BCI users and their needs. In J. R. Walpow & E. W. Walpow (Eds.), *Brain-computer interfaces: principles and practice* (pp. 317–323). Oxford: Oxford University Press.

Hochberg, L. R., Bacher, D., Jaroslewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., van der Smagt, P., & Donoghue, J. P. (2012b). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature, 485*, 372–375.

Hoffmann, S., & Falkenstein, M. (2011). Predictive information processing in the brain: Errors and response monitoring. *International Journal of Psychophysiology, 83*, 208–212.

Kitcher, P. (2011). *Science in a democratic society*. Amherst: Prometheus Books.

Laudan, L. (1978). *Progress and its problems: Toward a theory of scientific growth*. Berkeley: University of California Press.

Lucivero, F., & Tamburrini, G. (2008). Ethical monitoring of brain-machine interfaces, A note on personal identity and autonomy. *AI & Society, 22*, 449–460.

Mattia, D., Pichiorri, F., Molinari, M., & Rupp, R. (2012). Brain computer interface for hand motor function restoration and rehabilitation. In B. Z. Allison et al. (Eds.), *Towards practical brain-computer interfaces* (pp. 131–153). Berlin: Springer.

Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B., & Rosahl, S. (2007). *Intervening in the brain. Changing psyche and society*. Berlin: Springer.

Millán, J. del R., Ferrez, P. W., & Buttfield, A. (2006). *Non-invasive brain-computer interfaces* (*ESA, European Space Agency Report*: ID 05/6402).

Millán, J. del R., Renkens, F., Mouriño, J., & Gerstner, W. (2004). Brain-actuated interaction. *Artificial Intelligence, 159*, 241–259.

Millán, J. del R., Rupp, R., Müller-Putz, G. R., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kübler, A., Leeb, R., Neuper, C., Müller, K.-R., & Mattia, D. (2011). Combining BCI and assistive technologies: State-of-art and challenges. *Frontiers in Neuroscience, 4*, 161–168.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful modulation of brain activity in disorders of consciousness. *New England Journal of Medicine, 362*, 579–589.

Moran, D. (2010). Evolution of brain-computer interface: Action potentials, local field potentials and electrocorticograms. *Current Opinion in Neurobiology, 20*, 741–745.

Müller, K. R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods, 167*, 82–90.

Nehmzow, U. (2006). *Scientific methods in mobile robotics*. Berlin: Springer.

Nussbaum, M. C. (2006). *Frontiers of justice. Disability, nationality, species membership*. Cambridge: Harvard University Press.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science, 313*, 1402–1403.

Pichiorri, F., De VicoFallani, F., Cincotti, F., Babiloni, F., Molinari, M., Kleih, S. C., Neuper, C., Kübler, A., & Mattia, D. (2011). Sensorimotor rhythm-based brain-computer interface training: The impact on motor cortical responsiveness. *Journal of Neural Engineering, 8*. doi:10.1088/1741-2560/8/2/025020.

Proctor, R. (2008). Agnotology: A missing term to describe the cultural production of ignorance (and its study). In R. Proctor & L. Schiebinger (Eds.), *Agnotology: The making and unmaking of ignorance* (pp. 1–33). Stanford: Stanford University Press.

Rothaar, M. (2010). Human dignity and human rights in bioethics: The Kantian approach. *Medicine, Health Care and Philosophy, 13*, 251–257.

Santoro, M., Marino, D., & Tamburrini, G. (2008). Robots interacting with humans. From epistemic risk to responsibility. *Artificial Intelligence & Society, 22*, 301–314.

Schneider, M.-J., Fins, J. J., & Walpow, J. R. (2012). Ethical issues in BCI research. In J. R. Walpow & E. W. Walpow (Eds.), *Brain-computer interfaces: principles and practice* (pp. 373–383). Oxford: Oxford University Press.

Shih, J. J., Krusienski, D. J., & Wolpaw, J. R. (2012). Brain-computer interfaces in medicine. *Mayo Clinic Proceedings, 87*, 268–279.

Tamburrini, G. (2009). Brain to computer communication: Ethical perspectives on interaction models. *Neuroethics, 2*, 137–149.

Tamburrini, G., & Mattia, D. (2011). Disorders of consciousness and communication: Ethical motivations and communication-enabling attributes of consciousness. *Functional Neurology, 21*, 51–54.

United Nations. (1948). Universal declaration of human rights. http://www.un.org/en/documents/udhr/index.shtml

United Nations. (2006). Convention on the rights of persons with disabilities. http://untreaty.un.org/English/notpubl/IV_15_english.pdf

Várkuti, B., Guan, C., Pan, Y., Phua, K. S., Ang, K. K., Kuah, C. W., Chua, K., Ang, B. T., Birbaumer, N., & Sitaram, R. (2013). Resting state changes in functional connectivity correlate with movement recovery for BCI and robot-assisted upper-extremity training after stroke. *Neurorehabilitation and Neural Repair, 27*, 53–62.

Velliste, M., Perel, S., Chance, S. M., Whitford, A. S., & Schwartz, A. B. (2008). Cortical control of a prosthetic arm for self-feeding. *Nature, 453*, 1098–1101.

Vidaurre, C., & Blankertz, B. (2010). Towards a cure for BCI illiteracy. *Brain Topography, 23*, 194–198.

Vlek, R. J., Steines, D., Szibbo, D., Kübler, A., Schneider, M.-J., Haselager, P., & Nijboer, F. (2012). Ethical issues in brain-computer interface research, development, and dissemination. *Journal of Neurologic Physical Therapy, 36*, 94–99.

Wolpaw, J. R., & Walpow, E. W. (2012). *Brain-computer interfaces: Principles and practice*. Oxford: Oxford University Press.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Purtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113*, 767–791.

Attention Deficit Hyperactivity Disorder: Improving Performance Through Brain–Computer Interface

46

Imre Bárd and Ilina Singh

Contents

Abstract

Novel neurotechnologies, such as brain-computer interfaces (BCI), are generating significant scientific and popular interest. A certain BCI technology, neurofeedback (NF), is increasingly used for managing the symptoms of many conditions, most notably attention deficit hyperactivity disorder (ADHD). Although growing evidence suggests that the method is promising, there is no consensus in the scientific literature about its efficacy, and different sources offer contradictory evaluations. Although neurofeedback has received comparatively little scholarly attention from ethicists, it presents numerous dilemmas that warrant consideration. While the

I. Bárd (✉)
London School of Economics and Political Science, London, UK
e-mail: i.bard@lse.ac.uk

I. Singh
Department of Social Science, Health & Medicine, King's College London, London, UK
e-mail: ilina.singh@kcl.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_124,

method is already widely used and its acceptability can be expected to grow, the precise mechanism of action and possible adverse effects of neurofeedback are poorly understood at present. The current regulatory landscape of neurofeedback devices seems inadequate, and in particular, the growing commercialization of BCIs and lack of oversight over EEG-based toys and games present a challenge for neuroethical analysis. After a brief discussion of ADHD, and the emergence of neurofeedback, this chapter provides an overview of assessments of NF's efficacy and a brief survey of some of the ethical and social aspects of the method for pediatric ADHD. The questions covered include adverse effects, regulation, responsible communication, identity considerations, and the enhancement use of BCIs.

Introduction

Novel neurotechnologies, such as brain-computer interfaces (BCIs) and brain stimulation methods, are generating a lot of scientific research and increased media interest (Erp et al. 2012; Kennedy 2011). The ethical and social issues raised by some of these developments are continuously being addressed (Clausen 2008, 2009); however, questions surrounding the use of noninvasive devices, and specifically noninvasive BCIs, have received comparatively less scholarly attention. Unlike invasive BCI technologies that require complex medical interventions, such as neurosurgery, noninvasive devices are relatively easily accessible and can often be used by individuals without expert support. A medical diagnosis is not necessary for purchase or use of these devices, and they can easily become available for both medical and nonmedical purposes.

The use of a certain BCI technology, EEG-based biofeedback, also known as neurofeedback, is growing rapidly as a treatment for various conditions, including attention deficit hyperactivity disorder (ADHD), anxiety, depression and epilepsy (Birbaumer et al. 2009; Budzynski et al. 2009). Recently, some companies have brought neurofeedback software and equipment to the commercial market, which are advertised as improving the symptoms of ADHD, and numerous toys and video games are hitting the shelves that use largely the same technology.

This chapter will give an overview of the use of neurofeedback BCIs for ADHD and address some of the neuroethics questions associated with this practice. The first section will introduce relevant facts about ADHD, followed by an explanation of neurofeedback BCI and its development, how it is practiced, and the current scientific status of the method. The last section will suggest some of the key elements of an ethical framework for use of BCIs for ADHD, using neurofeedback as a current case to think with.

Attention Deficit Hyperactivity Disorder

Attention deficit hyperactivity disorder (ADHD) is among the most commonly diagnosed childhood psychiatric conditions worldwide. Its core symptoms are

hyperactivity, impulsiveness, and inattention, which give rise to three subtypes of ADHD: the predominantly inattentive, the predominantly hyperactive-impulsive, and the combined types. Affected children can have significant difficulties in coping with daily demands at home and in school and may experience stigmatization and exclusion (Singh 2012).

Currently, there are no reliable biomarkers or neuropsychological tests for ADHD; its diagnosis is therefore based on a careful examination of medical history to rule out other possible conditions and, most importantly, on behavior rating scales completed by parents and teachers to assess the nature and degree of symptomatology. The fact that diagnosis relies heavily on subjective appraisals of perceived behavior highlights the importance of social factors for the ways in which the condition is understood. Whatever is considered to be an "impairment" is also contingent upon cultural notions of what constitutes proper childhood behavior, pressures to succeed, and the broader social and educational context in which childhood behaviors are situated (Gualtieri and Johnson 2005).

Diagnostic manuals, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM), consider the diagnosis of ADHD to be categorical and binary – one either has it, or not. However, it is difficult to sharply distinguish normal and pathological symptoms, and recently, there have been calls to move towards spectrum-based approaches in the classification of mental disorders, which may be more appropriate in the case of ADHD as well (Singh 2008; Swanson et al. 2009; Marcus and Barry 2011).

The prevalence rate of ADHD is somewhat difficult to ascertain because numbers vary widely depending on measurement methodology, the sampled population, geographic location, and the diagnostic criteria used. According to one estimate the worldwide prevalence of ADHD is approximately 5 % (Polanczyk et al. 2007), while the CDC reports that in the United States 9.5 % of children aged 4–17 years have ADHD (CDC 2010). Furthermore, the diagnosis of the condition seems to be on the rise, with a reported 66 % increase in the US between 2000 and 2010 (Garfield et al. 2012).

Psychostimulant drugs are the most common form of treatment for ADHD. Their use dates back to the 1930s, and the most commonly used drug, methylphenidate (Ritalin), has been on the market since 1955 (Singh 2002a). The use of stimulants has been increasing steadily over the past decades, and it is estimated to be at 3.5 % overall among children and adolescents in the USA. Prescription stimulant use is highest among 6–12-year-olds, at 5.1 %, while use among adolescents is growing at the fastest rate and is also around 5 %. Boys are three times more likely to receive a diagnosis of ADHD than girls (Zuvekas and Vitiello 2012).

The effectiveness of psychostimulant medications has been established by numerous clinical trials; however, the NIMH-funded Multimodal Treatment Assessment Study (MTA) found that although drug treatment is associated with short-term benefits, these effects dissipate over time. Even the best available treatment, a carefully managed combination of behavioral interventions and psychostimulants, which is the optimal treatment recommended by the American Academy of Pediatrics (American Academy of Pediatrics 2011), provided inadequate benefit in close to

one third of the participants (Molina et al. 2009; Swanson et al. 2001). Furthermore, at follow-ups both at 6 and 8 years, there were almost no significant differences between randomized groups receiving psychostimulant, behavioral, or combined therapy for 14 months. However, with the exception of two variables (the Multidimensional Anxiety Scale for Children anxiety and driving accidents/citations), all treated groups scored significantly lower than the normative control group. Measured variables included behavioral ratings, academic achievement, and overall functioning assessed by the Columbia Impairment Rating Scale. Based on the MTA's findings, children with the least severe initial presentation of ADHD and those with sociodemographic advantage who respond best to any treatment have the best prognosis (Jensen et al. 2007; Molina et al. 2009).

Both the clinical reality of the ADHD diagnosis and the use of psychostimulant drugs in children are the subjects of a heated, ongoing debate, which takes place in the popular press as well as in academic journals. While some critics dismiss ADHD as inappropriate medicalization of normal childhood behavior, others advance more nuanced critiques that do not question the reality of the diagnosis entirely but articulate concerns about overdiagnosis and the wider social dimensions of children's behavior (see Singh 2008 for an overview).

Any intervention to manage ADHD has to be viewed against this complex background. ADHD diagnosis is, at best, an ambiguous medical label that captures a wide range of individuals whose level of impairment can differ significantly. Moreover, stimulant drug treatments have been shown to be effective both in individuals diagnosed with ADHD and in individuals who do not meet criteria for ADHD. This contributes a further opacity to the management of ADHD because the distinction between "treatment" of impairing symptoms and "enhancement" of cognitive capacities can be particularly subtle. Therefore, this chapter will discuss BCIs both as a treatment for ADHD and as an enhancement strategy for individuals who wish to improve their attentional capacities.

What Are BCIs?

Brain-computer interfaces (BCIs) translate readings of the electrical activity of the brain into commands that can be interpreted by computers. In essence, this allows users to operate software, devices, or prosthetics with nothing but their brain states or "thoughts." BCIs can be grouped into two categories. Invasive devices require a neurosurgical intervention to place electrodes on the surface of the brain or to implant them inside the brain. Noninvasive BCIs employ sensors, which are placed on the surface of the scalp to record electroencephalographic (EEG) signals. The primary medical application of BCIs is within the field of assistive technologies for patients with locked-in syndrome and other disabilities, where the method is used to restore lost functionality, such as vision, speech, or control over prosthetics or a wheelchair (Berger et al. 2008).

Irrespective of the level of invasiveness, there are two different approaches to BCIs, which can be distinguished, on the basis of which party is doing most of the

learning: the brain or the computer, although most technologies employ both to some extent. In the *machine learning* approach, sophisticated learning algorithms adapt to the individual user's neural activity through a process of calibration as the subject performs certain tasks, such as imagining body movements. In the *biofeedback* approach, subjects learn to produce certain desired brain states with the help of visual and auditory feedback about their real-time neural activity (Dornhege et al. 2007). Beyond assistive technologies, noninvasive BCIs are increasingly used for a wide range of purposes, including coma detection, meditation training, computer usability research, alternative therapies, performance enhancement, and gaming (Desney and Nijholt 2010).

Most of the research on BCIs to date has been conducted with EEG-based devices. Although there is emerging research about fMRI neurofeedback for ADHD, this technology is still highly experimental (http://public.ukcrn.org.uk/search/StudyDetail.aspx?StudyID=12668). In the future, hybrid devices might appear that combine measurements of various physiological responses, such as electrodermal activity or heart rate, with BCIs.

Because it is the most widely used and studied method, the rest of the chapter will focus on the use of EEG neurofeedback for pediatric ADHD. While the method can look back at a history of several decades, the precise mechanisms of action are still poorly understood.

Neurofeedback

A Brief History of Neurofeedback

At the end of the eighteenth century, the work of Luigi Galvani and Giovanni Aldini introduced the idea of electricity as the agent of peripheral nerve function. This was a revolutionary discovery and in a way it also signaled the start of psychiatry's lasting relationship with the electrical properties of the brain (Ochs 2004). About 100 years later, in 1875, Richard Caton conducted the first electrophysiological recordings of the central nervous systems of animals. Caton was interested in cortical localization, and he postulated that there must be some kind of relationship between the weak currents he had measured and specific brain functions. His work was mostly forgotten until 1929, when Hans Berger, the inventor of the modern EEG, acknowledged Caton's essential role in researching the electrical activity of the brain (Finger 1994).

Berger was the first to record brain waves on paper thus creating the electroencephalogram. Through his work on humans Berger found that the EEG consisted of certain rhythms of which he identified two distinct waves, alpha and beta, and he also postulated that the EEG could be used to detect clinical pathologies (Borck 2005). Throughout the 1930s, the EEG quickly became associated with the promise of deciphering the language of the brain as it gives rise to the mind, and its visual representation, the lines of ink on paper came to be seen as the bridge between biological processes and mental life.

The early history of the EEG intersects with the early history of ADHD. In 1937, Charles Bradley conducted the first trial of a psychostimulant drug, Benzedrine, for the treatment of "behavior problem children" (Bradley 1937). This was a broad category at the time, and it included children who would likely be classed as ADHD today (Singh 2002b). A year later, Bradley, along with Herbert Jasper and Philip Solomon, conducted the first EEG examinations of behavior problem children (Jasper et al. 1938). In this and in subsequent experiments (Lindsley and Cutts 1940), they found an increased rate of slow waves compared to normal controls. However, it is important to note that at the time it was difficult to interpret EEG measurements because there were no normative databases available to compare findings to, and advanced methods for quantifying EEG data did not yet exist. The 1960s brought a number of significant breakthroughs. First, the advent of sophisticated computers led to the emergence of quantitative EEG (qEEG), which gradually enabled more and more features of EEG data to be analyzed, such as frequency, amplitude, morphology, and symmetry (Cantor 1999). In addition, researchers began to establish normative databases that allowed individual measurements to be compared to a representative pool of data (Thatcher and Lubar 2009).

The notion of neurofeedback is predicated on the idea of operant conditioning developed by the father of behaviorism Burrhus Frederic Skinner in the 1930s. It is a form learning where the occurrence of a certain behavioral response is reinforced or extinguished through rewards or punishments (Skinner 1938, 1963). With the emergence of cybernetics, the notion of feedback gained central importance in understanding biological organisms. In neurofeedback, the ideas of the operant conditioning of behavior and feedback in self-regulation were brought together (Kropotov 2009). Thus, neurofeedback is "an operant conditioning procedure whereby an individual modifies the amplitude, frequency, or coherency of the electrophysiological dynamics of his/her own brain" (Thatcher 1999, p. 29). Dr. Joe Kamiya was the first to investigate whether there was any correlation between 1st person subjective experience and certain EEG patterns, namely, bursts of alpha (Kamiya 1968). He also thought to connect an EEG to a machine that sounded a tone whenever subjects produced alpha bursts (Nowlis and Kamiya 1970). He found that such bursts were associated with subtle cues and that through introspection and feedback of their own brain states participants could learn to control alpha consciously. Barry Sterman used an operant conditioning paradigm and trained cats to increase a certain band of the beta wave in the 12–15 Hz range known as the sensorimotor rhythm (SMR), which is associated with immobility (Sterman et al. 1969). It was later discovered accidentally in a NASA study that this training had made the cats highly seizure resistant. Subsequently, this discovery opened the prospect for medical applications of the method in the treatment of epilepsy. These findings generated a lot of research interest and by the early 1970s led to the birth of the discipline of biofeedback, which was also strongly supported in the US during the Vietnam War to produce methods of performance enhancement, especially under elevated stress (Budzynski 1999).

Drawing on successes in the suppression of seizures, Lubar and colleagues hypothesized that training SMR could be used effectively for the reduction of

hyperkinetic symptoms in children and they produced impressive results (Lubar and Shouse 1976; Shouse and Lubar 1979). By this time, stimulant medications were available treatments for hyperkinetic children, and the higher rate of slow waves characteristic of these children has been used as a predictor of positive response to stimulant drugs (Satterfield et al. 1973). As more and more evidence showed that children with ADHD showed cortical slowing in frontal regions, Lubar and Lubar performed the first experiment using a beta enhancement/theta suppression paradigm, which became the basis of most subsequent neurofeedback procedures for ADHD (Lubar and Lubar 1984).

After initial successes and great expectations, the field of neurofeedback suffered a critical setback and fell into disrepute, the reasons for which are still disputed. The method gradually became associated with unrealistic claims related to consciousness expansion, which were considered fringe at the time. Numerous studies reported on lower therapeutic efficacy than had been expected, and neurofeedback was also perceived by many to be dangerously close to mind control. However, the 1990s and early 2000s brought a renewed interest in the method as computerized EEG became more easily accessible, and research in brain science was revitalized by numerous breakthroughs and the "Decade of the Brain" (Evans and Abarbanel 1999).

Neurofeedback for ADHD Today

Today neurofeedback is growing rapidly as a form of alternative treatment, marketed as a highly effective intervention for both therapy and personal growth (Demos 2005). There are two primary forms of receiving neurofeedback: either with a trained professional or by using commercial devices at home. Professional neurofeedback therapists usually perform an extensive quantitative EEG (qEEG) measurement at the first session, and the client's qEEG data is compared to a normative database in order to determine the specific training paradigm and electrode placement that will be used. Normative qEEG databases, such as the Neurometric Analysis System, have received FDA clearance (John et al. 1983), but they can only serve as aids to diagnosis, treatment planning, and follow-up. Recent reviews argue for a more prominent clinical role of qEEG in child and adolescent psychiatry (Chabot et al. 2005; Hirshberg et al. 2005). Neurofeedback therapists also produce a colorful "brain map" to visualize individual EEG characteristics, and ethnographic evidence shows that clients often find this visualization as well as the brain-based explanation of their symptoms highly convincing and compelling (Brenninkmeijer 2010).

ADHD in particular is understood to be related to under-arousal in frontal areas and cortical slowing, which means that children show an excessive amount of slow waves and a reduction in faster waves, which are associated with alertness and concentration (Monastra et al. 1999). Therefore the most common neurofeedback training method is to suppress slow, theta activity while increasing faster waves in the beta range. Various protocols have been developed that draw on this principle

(Monastra 2005), but recently researchers have used slow cortical potential (SCP) training paradigms as well (e.g., Heinrich et al. 2004). SCPs are event-related changes in cortical polarization that play a role in the preparatory distribution of attentional resources (Birbaumer 1999; Strehl et al. 2006). Other than theta/beta training, which aims at changing abnormalities in resting EEG and influencing the ratio of slow waves to faster waves in certain areas, SCP training aims at increasing cortical negativity and thereby improving cognitive control (Heinrich et al. 2004).

Considerable differences exist between neurofeedback systems in terms of the feedback they give. The simplest type provides a visual representation of the subject's EEG activity and sounds a tone when the desired patterns are produced. Others can be linked to a video or DVD player where the image size and resolution increase or decrease based on EEG activity. Go/No-Go systems employ simple video games, such as Pac-Man, where the player's character moves if appropriate waves are produced but freezes when the subject is not in the desired state. Still other systems can be connected to commercially available video games using Sony PlayStation or Nintendo X-Box consoles where the player's video game controller loses sensitivity or otherwise impedes gameplay if the subject is not in the expected brain state (http://www.smartbraintech.com). Most recently, researchers have begun to combine neurofeedback technology with immersive virtual reality environments (http://openvibe.inria.fr/a-bci-based-virtual-reality-solution-for-adhd-treatment-prototype/).

The FDA's 2013 decision to permit marketing of the first ever EEG diagnostic test for ADHD has received mixed reactions from experts. The system in question measures theta-beta ratio and produces a report to aid clinicians' evaluations but it is not intended to be a stand-alone method in diagnosis (http://www.fda.gov/newsevents/newsroom/pressannouncements/ucm360811.htm).

BCI games have also entered the commercial market (Cole 2007), and some of these games are advertised as effective tools for improving the symptoms of ADHD (e.g., FocusPocus, SmartBrain). Although neurofeedback professionals warn against unsupervised and one-size-fits-all use of the method (Hammond 2011), the practice of home neurofeedback is likely to grow, given that the method combines gaming with an ostensible therapeutic function. This may encourage both parents and children hoping for an effective alternative to drug treatment for ADHD (Fuchs et al. 2003).

Key Elements of an Ethical Framework: BCIS for ADHD

The following sections will discuss two foundational ethical issues – safety and efficacy of BCI technologies – and analyze four practical implications of these concerns: regulation and commercialization, responsible communication, identity considerations, and the enhancement use of BCIs. This is not, of course, an exhaustive presentation of ethical concerns that arise from BCI technologies. The focus on safety and efficacy is meant to provide the conceptual foundations for an ethical framework for the use of BCIs for ADHD in children. Considering that BCI

technologies for ADHD are still at an early stage of development, and are not widely used, it is unwise to indulge in a speculative ethics and to develop an account, which has little basis in evidence or in reality.

How Effective Is Neurofeedback?

Research on neurofeedback has been growing quickly since the mid-1990s, although the precise function of brain rhythms is still poorly understood (Frank 2009). Different sources offer differing accounts of the level of scientific support in favor of neurofeedback in the treatment of ADHD. While there is a considerable body of case studies that demonstrates equal or superior effectiveness compared to stimulant drugs, there are very few controlled and methodologically rigorous studies with large sample sizes. Furthermore, there is an ongoing debate in the literature about how the available scientific data should be evaluated and interpreted. In particular, reviewers disagree on the methodological robustness of available findings, and some critics question whether the observed positive effects of neurofeedback are really caused by EEG training and not other factors. In the past decade, the efficacy of neurofeedback for the treatment of ADHD has been reviewed on numerous occasions, with the various studies arriving at very different conclusions. The reviews and analyses most commonly refer to standardized systems of evaluation, such as the five-point scale of the American Psychological Association, used for the categorization of scientific evidence in support of different treatments, which ranges from “Not Empirically Supported” (Level 1) to “Efficacious and Specific” (Level 5).

Monastra and colleagues conducted the first review of case studies and controlled-group studies of neurofeedback in the treatment of ADHD. Although both case studies and controlled studies demonstrated beneficial effects of neurofeedback on the core symptoms of ADHD, the authors pointed out that studies had not been adequately blinded and thus could not control for expectancy and other nonspecific treatment effects. Still, the paper concluded that neurofeedback could be considered a Level 3 “probably efficacious” intervention for ADHD (Monastra et al. 2005). The same year, Loo and Barkley published a more critical evaluation of the method, in which they stated that neurofeedback studies for ADHD were either methodologically weak or they could not demonstrate superior efficacy compared to placebo or no-treatment control (Loo and Barkley 2005). The National Resource Center on AD/HD also published a non peer-reviewed online paper discussing eight major controlled studies that had been conducted through 2005 and concluded that neurofeedback for ADHD was a Level 2 “possibly efficacious” method (WWK 2008). The report also stressed the lack of proper randomization and inadequate blinding as the most important limitations of available studies while pointing out that some evidence suggests the use of neurofeedback for ADHD in schools could be associated with greater class inclusion and financial savings.

Hodgson and colleagues conducted a meta-analytic review of nonpharmacological treatments for ADHD, which included behavior modification, neurofeedback therapy, multimodal psychosocial treatment, school-based programs,

working memory training, parent training, and self-monitoring. Neurofeedback produced the largest average weighted effect size on outcome measures, and the authors concluded that it was the most efficacious evidence-based nonpharmacological intervention for reducing ADHD symptoms (Hodgson et al. 2012). However, this paper received criticism for neglecting recent research findings and for the small number of studies it included (Rothenberger and Rothenberger 2012).

In 2009, Arns and colleagues conducted a thorough meta-analysis, which included 15 studies and thus drew on a larger pool of findings than the earlier reviews. The authors came to the conclusion that neurofeedback for ADHD warranted the highest possible, Level 5 classification. Of the core symptoms of ADHD, Arns et al. reported high effect sizes of neurofeedback for inattention and impulsivity and medium effect size for hyperactivity (Arns et al. 2009). The meta-analysis looked at prospective controlled studies, prospective pre-/post-design studies and retrospective pre-/post-design studies, although they considered randomized and nonrandomized investigations together, which makes it difficult to draw solid conclusions. Although the authors acknowledged some methodological limitations that had been pointed out by earlier reviews, they still considered the available evidence strong enough to support neurofeedback as an “efficacious and specific” intervention. Understandably, neurofeedback service providers predominantly refer to this most favorable evaluation of the method.

An extensive review of studies through September 2010 by Lofthouse and colleagues articulated firm disagreement with the Arns et al. paper and claimed that neurofeedback for pediatric ADHD could only be considered “probably efficacious” and that a large, multisite, double-blind randomized control trial was necessary to settle the question concerning efficacy (Lofthouse et al. 2012a). The authors’ central criticism was that on the basis of current evidence it is not possible to establish whether the observed benefits are due to specific effects of neurofeedback or other nonspecific ones, such as “selection effects, participant history, regression to the mean, placebo response, maturation, practice with assessment measures, and/or participant–rater–experimenter expectancies” (Lofthouse et al. 2012a, p. 15). They also pointed out that studies did not sufficiently monitor the use of concomitant treatments that might have influenced outcomes and there was no reliable follow-up data available. Researchers also failed to report on any observed adverse effects, and there was considerable variability in the administration of neurofeedback both in terms of the specific systems and software used and in terms of treatment length and frequency (Lofthouse et al. 2012a).

In a subsequent review of studies since 2010 Lofthouse et al. acknowledged positive persistent benefits associated with neurofeedback and medium-to-large improvement. However, methodological concerns prompted the authors to maintain the position that available evidence did not support the time and money parents would have to invest into a course of neurofeedback treatment (Lofthouse et al. 2012b).

The possibility of proper blinding and sham control is of particular importance, and there has been disagreement in the literature about whether it is possible and/or ethical at all (Lofthouse et al. 2012a; Leins et al. 2007; La Vaque and Rossiter 2001), given that wait-list and placebo condition studies lasting several tens of

session are in conflict with the Declaration of Helsinki (World Medical Association 2000). An NIMH-funded feasibility study with 39 participants conducted by Arnold and colleagues found that proper blinding was indeed possible, that a sham-condition did not lead to high drop-out rates and that the optimal treatment frequency was 3×/week (Arnold et al. 2013). But the question concerning the feasibility of sham control is still unresolved. A true sham control should be inert, and in the case of neurofeedback, researchers have provided subjects with random feedback instead of feedback contingent on their EEG activity. However, the system does reinforce some activity it can still occur that it reinforces the desired EEG activity. Furthermore, the regular participation in a treatment setting as well as the repeated effort to pay attention during session might also produce some effects. A further complication is that all studies to date that used sham feedback found that both the active and control groups improved in primary ADHD symptoms with no significant differences between the two conditions, suggesting that the method is not better than placebo or that the main treatment effects are nonspecific. For such reasons, Loo and Makeig do not recommend the most commonly used theta/beta neurofeedback neither as a first-line nor as an adjunct treatment until there is convincing evidence that the method is superior to placebo or equally effective as established treatments. However, the authors note that SCP neurofeedback may prove effective as an adjunct therapy in a significant subset of children, but more research is needed to identify which methods are effective in which ADHD patient populations (Loo and Makeig 2012).

A further, complicating issue concerns the reliability of neurofeedback technology itself. Some studies suggest that EEG-based BCI technologies are limited by the spatial resolution at which brain signals are measured and they are also vulnerable to interference from facial muscles or blinking, movements that are all the more difficult to control by children (Zander and Kothe 2011).

In summary, there is a wide range of opinions about neurofeedback and different appraisals of the available studies. However, at this time, a majority of reviews are cautious about recommending neurofeedback for treatment of ADHD. A group of researchers comprising both neurofeedback experts and ADHD treatment outcome investigators is preparing a multisite double-blind sham-controlled study with follow-up assessments that might settle the question of efficacy (Lofthouse et al. 2012b).

Adverse Effects

While neurofeedback research can look back at a history of several decades, there is very little research about the long-term effects of the method and especially on its impact on developing brains. Although there are no reports of adverse side effects in the literature, and neurofeedback practitioners stress the methods' safety, there is no systematic evaluation of this issue (Rutger et al. 2012). There are several points in relation to possible adverse effects with respect to neurofeedback training for ADHD in particular.

First, more research is needed on what we might call the “ecology of brain waves”. For example, some research suggests that theta waves, which neurofeedback training often tries to decrease, play an important role in musical performance (Egner and Gruzelier 2003). Thus, while increasing or decreasing the rate of certain frequencies or changing their ratios, such as in the case of theta/beta, might prove beneficial from one perspective, it is not clear how that change could influence other capacities and activities (Baum 2011).

Second, many EEG biofeedback-based toys available on the commercial market today reinforce the generation of theta waves in order to win (such as Mindflex and Mindball). As already mentioned, children with ADHD already show elevated levels of theta, which may suggest that certain BCI games could actually exacerbate their symptoms.

Moreover, neurofeedback systems, which are attached to computer or video games (e.g., SmartBrain, Focus Pocus) raise an interesting question about the relationship between screen time and ADHD. Some research suggests that there may be an adverse connection between internet- and video-game use, and ADHD, such that the risk of addictive gaming behavior may be elevated and ADHD symptoms worsened (Weiss et al. 2011). The connection between gaming and ADHD needs more rigorous scientific evaluation to better assess the relative risks and benefits of this particular mode of intervention.

The repetitive use of certain brain pathways through EEG-type devices has unknown consequences for brain structure and function. The brain has been shown to be malleable through indirect influences such as learning; therefore, it is likely that the brain’s plasticity will be impacted by high use of interventions such as neurofeedback. Further research is necessary to better understand the impact of noninvasive BCI interventions over time and how these interact with the processes of brain development and maturation in children (Hildt 2010).

Finally, the Nuffield Council on Bioethics (NCOB 2013) notes that one of the most insidious risks of BCI technologies such as neurofeedback may be the psychosocial effects of disappointing results following high expectations. This concern is compounded by the relatively unregulated space that noninvasive BCIs inhabit. As the following section outlines, the lack of regulation has implications for both the marketing and the delivery of neurofeedback systems for ADHD, as well as, of course, consumer safety and autonomy.

Implications of Safety and Efficacy Considerations

Regulation and Commercialization

Commercialization of BCI devices for ADHD as well as for nonmental health conditions is currently running ahead of regulation strategies to manage the uncertainty surrounding these devices. In the case of neurofeedback, practitioners stress that the method needs to be administered by a trained professional and that it is definitely not a “kid’s toy” (Thatcher 1999). However, neurofeedback

is becoming exactly that as more and more companies are bringing neurofeedback systems for home use to the commercial market with almost no regulatory oversight.

In the United States, neurofeedback devices are considered Class II medical devices by the Food and Drug Administration (FDA). This is the middle category on the FDA's three level system of classification, which is based on potential risks and harms. The FDA has cleared the use of neurofeedback devices in 1976 for relaxation and muscle reeducation only. A specific type of device, namely, a "prescription battery powered device that is indicated for relaxation training and muscle reeducation" (http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=882.5050) was granted 510 k exempt status, which means that manufacturers of such devices do not have to file a (510 k) Premarket Notification to the FDA before bringing their product to the market. Manufacturers are still required to register with the FDA, and any device that is different from the above still has to obtain FDA clearance. The FDA has accepted Premarket Notification for a number of EEG biofeedback devices, which have been deemed "substantially equivalent" to those marketed in 1976 or before. This means that neurofeedback cannot be promoted as a treatment for conditions like ADHD, but professionals can use the devices as part of their practice for numerous off-label purposes. However, if any new device received FDA clearance with an indication for ADHD, then all other devices deemed substantially equivalent would receive the same clearance as well, although considerable differences exist between the various systems in terms of their technical features, software, and training paradigms.

The European Union has three Directives that regulate medical devices, of which one is relevant to noninvasive BCI technologies. The underlying objectives of these Directives are to ensure safety and efficacy of the devices, as well as to remove barriers to trade within the EU. Companies must demonstrate that a device meets the requirements of a given Directive to receive the CE-mark, which is widely held to be a guarantee of quality, although it is not.

In the USA and the EU, medical devices can only be sold to licensed health care practitioners, and their sale must be reported to relevant national "competent authorities" such as the FDA and, in the UK, the Medicines and Healthcare Products Regulatory Authority (MHRA). However, an increasing number of noninvasive BCI technologies are delivered via a nonmedical route. Neurofeedback BCI devices are available to private individuals, numerous toys and video games that draw on similar technology are being brought to market, and many more can be expected in the near future (Nijboer et al. 2011). Because these are commercial products not marketed for medical purposes, they are not regulated by the FDA or the EU Directives. Direct-to-consumer (DTC) advertising is not covered by any of the EU Directives, nor does the FDA regulation on DTC advertising cover nonmedical devices.

It is unclear which range of private individuals and professionals will choose to use noninvasive BCI technologies. It is also unclear who the consumers of such technologies will be. Yet "brain training" aimed particularly at individuals with memory and attention difficulties is offered to children and families via private

clinics and on the Internet. The educational sphere may also be impacted: on the Internet, toys are sometimes marketed as tools that schools can use to improve the performance of easily distracted children (see http://www.vivifeye.com/) without firm scientific support for such claims (Foks 2005; Plischke et al. 2011).

Without adequate oversight and evidence the use of BCI technologies for ADHD might carry certain risks, especially if administration is undertaken in the absence of trained professional supervision. In the case of neurofeedback, there are no established criteria for what constitutes appropriate and inappropriate use and what kind of education and training is necessary to provide neurofeedback. At present, health care professionals can offer neurofeedback as part of their services without any special training. Although professional organizations like the Association for Applied Psychophysiology and Biofeedback (AAPB) offer certification, it is not required to practice. For this reason, some have argued that neurofeedback should ideally be limited to the clinical domain and practiced only by physicians and clinical psychologists (Giordano and DuRousseau 2011). Furthermore, neurofeedback professionals' guidelines firmly stress the importance of purchasing FDA-cleared equipment and using these within the scope of the individual's license; however, not all professionals abide by these rules (Striefel 2009).

A precautionary view considers the increasing commercialization of BCI technologies like neurofeedback to be problematic both from an ethical and a scientific perspective. There is limited understanding of the precise mechanisms through which different technologies exert effects, and effects on the developing brain are largely unknown. Given that considerable differences exist between various neurofeedback systems, mechanisms, and effects, these have to be defined more clearly in order to achieve differential regulation. This position urges a precautionary approach as well as stronger governmental oversight and mechanism to evaluate claims about safety and efficacy (Tamburrini 2009; Plischke et al. 2011).

The UK Nuffield Council on Bioethics' Report on Novel Neurotechnologies is the first to outline a substantive approach to the regulation of noninvasive BCIs and other novel neurotechnologies (forthcoming). The approach draws on an ethical framework that incorporates the precautionary principle but also recognizes the need for innovation in the area of novel neurotechnologies. The Nuffield framework seeks to balance the uncertainty surrounding current BCI technologies and the real need of patients suffering from complex, chronic brain diseases that cause significant mental, physical, and emotional suffering to individuals and their families. Many of the recommendations in the report involve soft regulation, in an effort to encourage both caution and innovation in the development and application of BCI technologies. Further discussion and debate are now necessary to achieve a sound policy approach to the regulation of BCI technologies; eventually more oversight and evaluation are certainly necessary (van Est et al. 2010). Strict guidelines, policies, and laws that are informed by neuroethical analyses and based upon neuroscientific information are required to guide the dissemination of neurofeedback-based devices in the public sphere (Giordano and Olds 2010). However, in the case of neurofeedback, this neuroscientifically informed

neuroethical analysis is complicated by the fact that currently no scientific consensus exists about the benefits of neurofeedback training. This lack of consensus is all the more pressing because consumers are already exposed to various, sometimes contradictory and often exaggerated claims about the method, which highlights the importance of responsible communication.

Responsible Communication and Neuroliteracy

Unlike psychostimulant drug therapy for ADHD, which can only be prescribed by medical doctors, private practitioners offer neurofeedback services, and BCI devices are already available commercially. Consumers of both professional neurofeedback training and commercial EEG-based products are therefore exposed to various claims about the method's efficacy and safety. A course of neurofeedback training can be a significant financial commitment over a lengthy period of time that despite hope-inspiring claims may not produce the desired outcomes (Lofthouse et al. 2012a). These factors foreground the relevance of responsible communication about and strict evaluation of novel neurotechnologies. Some argue that it may be ethically unsound to "market brainwave interface products as toys, games, and educational aides, without consideration of the possible, if not likely, consequences of putting BMNs [brain machine neurotechnologies] into the hands of consumers who do not realize the actual capabilities, as well as risks and harms, of these products" (Plischke et al. 2011, p. 224). BCI researchers have also argued that ethically sound, responsible communication and appropriately balanced media reporting are crucial (Nijboer et al. 2011).

Furthermore, although neurofeedback practitioners are expected to provide accurate and credible information about the method to their clients, and to continuously educate themselves about scientific developments (Striefel 2009), heated disputes surrounding the method make it rather difficult to achieve this clarity. In fact, service providers often cite only the most favorable reviews about the method. For example, the Association for Applied Psychophysiology and Biofeedback (AAPB), the first professional organization dedicated to biofeedback, does not list any research articles dated 2009 or later on its publicly available website, although several studies and reviews have been published since then. Furthermore an AAPB report from 2004 includes a long list of conditions alongside the level of biofeedback's clinical efficacy for that condition. This list includes ADHD and claims that EEG biofeedback is a Level 4, "efficacious intervention" (Yucha and Gilbert 2004) despite intense debates about such claims.

While stricter guidelines for neurofeedback professionals and device manufacturers are certainly necessary, the wider public may also need to become more neuro-literate. It is becoming increasingly important to gain basic familiarity with neuroscience, because as parents, teachers, and professionals, individuals confront arguments clad in neuroscientific terms for and against various technologies, products, therapies, educational approaches, and a whole host of other things. As neuroscience gains an ever more prominent role in shaping our understanding of the most diverse aspects of the human condition, "neuroliteracy" on the part of the wider public would contribute greatly to making more informed choices and decisions in the vast sea of claims and promises (Farah 2011).

Identity and Brain Interventions

An important consideration that flows from the discussion of efficacy of BCI technology for ADHD is the potential implications for identity. The ethical discussion of identity, intervention, and ADHD has taken place largely in relation to stimulant drug treatments. Here, the developmental dimensions of BCI technologies become particularly acute, as there is a substantive difference in ethical concerns between children and young people on the one hand and adults on the other.

Treatment of ADHD with stimulant medications has given rise to a passionate debate about the effect of such drugs on children's developing sense of self and moral identity. Critics have argued that children are drugged into obedience and conformity, which thwarts their moral development. Yet, recent research both from the social and natural sciences says that at least some of these fears are unfounded (Singh 2013; Hyman 2013). In fact, many children feel empowered to behave as moral agents and to better meet the expectations of their social environments with medication. Of course, this does not mean that the practice of treating children with psychostimulants is entirely unproblematic (Caplan 2013).

Like psychostimulants, BCI technologies explicitly aim to affect neural mechanisms putatively correlated with the capacity for cognitive self-control. Although neurofeedback is often portrayed as a way of learning to take control over brain activity, it is not entirely clear how much conscious effort is needed on the part of the child who is undergoing neurofeedback training, which is based on the principles of operant conditioning. In fact, some ethnographic evidence suggests that subjects need to engage in minimal effort and should remain passive observers of the process of retraining their brains. The literature often reports of the transformative power of the method, as it restores selves thought lost to disease or transforms one's self as symptoms of disorders improve (Brenninkmeijer 2010).

While drug therapy has provoked intense discussions and polarized opinions, neurofeedback is sometimes discussed as "crackpot charlatanism" and sometimes as a panacea (Ellison 2010), but it has not been addressed as an intervention that would somehow interfere with moral capacity and self-governance. This is somewhat remarkable, considering the fact that both methods operate directly on the brain to bring about potentially transformative effects. Neurofeedback thus provides an interesting opportunity to study the factors that might give rise to moral concern about a particular intervention. Are drug-based interventions more prone to criticism and moral concern than novel neurotechnologies? If so, why? Such questions cannot be answered on the basis of conceptual analysis alone and are in need of empirical evidence.

Enhancement

As already argued, the lack of biomarkers for diagnosing ADHD blurs the boundary between the treatment of symptoms and the enhancement of cognitive capacities. Performance enhancement and peak performance training have been important applications of neurofeedback since the 1970s. According to some evidence, the method can be used to improve artistic, cognitive, and sports performance, although efficacy is still debated and needs more thorough investigation (Vernon 2005;

Gruzelier et al. 2006; Keizer et al. 2010). The increasing commercialization of the method (see http://www.emotiv.com) and online platforms, such as OpenEEG, (http://openeeg.sourceforge.net) further complicate the ethical evaluation of neurofeedback for performance enhancement. The method's broad availability and the likely demand for performance enhancement put an increased emphasis on the need for careful scientific evaluation of adverse and long-term effects and highlight the need for ongoing societal deliberation about maximizing the benefits and minimizing the risks of novel neurotechnologies.

Conclusion

Although neurofeedback has been around for many decades, it is currently undergoing an explosion in both popular and scientific interest. Several companies are racing to bring new BCI products to the market in sectors ranging from medical diagnostics and assistive solutions, through gaming and entertainment, to neuromarketing. This trend is very likely to continue into the future as the technology becomes cheaper, more sophisticated and readily available. Open-source hardware and software projects contribute a lot to making BCIs more widely accessible. However, the jury is still out on whether neurofeedback is an effective and specific method of improving performance in ADHD in particular. The scientific literature is inconclusive and there is little understanding of both the mechanisms of action and possible adverse effects. Contradictory evaluations of available findings make it difficult to arrive at any solid conclusions, which is especially problematic given the method's increasing commercial availability in the form of toys, games, and educational aids. These are often aimed specifically at children with ADHD. These uncertainties highlight the importance of further targeted research and responsible communication about the method. They also bring to the fore the growing need for the wider public to gain a basic understanding of neuroscience to make informed decisions about available interventions. Furthermore, some aspects of the current medical device regulatory framework and the lack of oversight over commercial products are problematic. As numerous authors have argued, straightforward policies, based on neuroethical analyses and informed by neuroscientific knowledge, are urgently needed. Neurofeedback's easy availability might make it less apparent that a wide range of questions and concerns surrounding the method are still unresolved, which underlines the importance of a broader engagement with the topic.

Cross-References

- ▶ Brain–Machine Interfaces for Communication in Complete Paralysis: Ethical Implications and Challenges
- ▶ Developmental Neuroethics
- ▶ Ethical Implications of Brain–Computer Interfacing
- ▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes

▶ Human Brain Research and Ethics
▶ Impact of Brain Interventions on Personal Identity
▶ Mind, Brain, and Education: A Discussion of Practical, Conceptual, and Ethical Issues
▶ Neuroenhancement
▶ Neuroethical Issues in the Diagnosis and Treatment of Children with Mood and Behavioral Disturbances
▶ Neuroscience, Neuroethics, and the Media
▶ Normal Brain Development and Child/Adolescent Policy
▶ Research in Neuroenhancement

References

American Academy of Pediatrics. (2011). ADHD: Clinical Practice Guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics, 128*, 1007–1022. doi:10.1542/peds.2011-2107B.

Arnold, L. E., Lofthouse, N., Hersch, S., Pan, X., Hurt, E., Bates, B., Kassouf, K., Moone, S., & Grantier, C. (2013). EEG neurofeedback for ADHD: Double-blind sham-controlled randomized pilot feasibility trial. *Journal of Attention Disorders, 17*, 410–419 [first published on May 22, 2012].

Arns, M., de Ridder, S., Strehl, U., Breteler, M., & Coenen, A. (2009). Efficacy of neurofeedback treatment in ADHD: The effects on inattention, impulsivity and hyperactivity: A meta-analysis. *Clinical EEG and Neuroscience, 40*(3), 180–189.

Baum, M. (2011). "Focus Pocus" and beyond: Consumer brain computer interfaces for health, self-improvement and fun. Practical Ethics Blog, University of Oxford http://blog.practicalethics.ox.ac.uk/2011/08/%E2%80%9Cfocus-pocus%E2%80%9D-and-beyond-consumer-brain-computer-interfaces-for-health-self-improvement-and-fun/

Berger, T. W., Chapin, J. K., Gerhardt, G. A., McFarland, D. J., Principe, J. C., Soussou, W. V., Taylor, D. M., & Tresco, P. A. (2008). *Brain-Computer interfaces: An international assessment of research and development trends*. Dordrecht: Springer.

Birbaumer, N. (1999). Slow cortical potentials: Plasticity, operant control, and behavioral effects. *The Neuroscientist, 5*, 74–78.

Birbaumer, N., Ramos Muguialday, A., Weber, C., & Montoya, P. (2009). Neurofeedback and brain-computer interface clinical applications. *International Review of Neurobiology, 86*, 107–117.

Borck, C. (2005). *Hirnströme. EineKulturgeschichte der elektroenzephalographie*. Göttingen: Wallstein.

Bradley, C. (1937). The behavior of children receiving benzedrine. *The American Journal of Psychiatry, 94*, 577–585.

Brenninkmeijer, J. (2010). Taking care of one's brain: How manipulating the brain changes people's selves. *History of the Human Sciences, 23*(1), 107–126.

Budzynski, T. H. (1999). From EEG to neurofeedback. In J. R. Evans & A. Abarbanel (Eds.), *Introduction to quantitative EEG and neurofeedback*. San Diego: Academic.

Budzynski, T. H., Budzynski, H. K., Evans, J. R., & Abarbanel, A. (Eds.). (2009). *Introduction to quantitative EEG and neurofeedback*. San Diego: Academic Press.

Cantor, D. S. (1999). An overview of quantitative EEG and its applications to neurofeedback. In J. R. Evans & A. Abarbanel (Eds.), *Introduction to quantitative EEG and neurofeedback*. San Diego: Academic.

Caplan, A. (2013). Accepting a helping hand can be the right thing to do. Journal of Medical Ethics, 39, 367–368. doi:10.1136/medethics-2012-100879.

Centers for Disease Control and Prevention. (2010). Increasing prevalence of parent-reported attention deficit/hyperactivity disorder among children: United States, 2003 and 2007. *MMWR. Morbidity and Mortality Weekly Report, 59*, 1439–1443.

Chabot, R. J., di Michele, F., & Prichep, L. (2005). The role of quantitative electroencephalography in child and adolescent psychiatric disorders. *Child and Adolescent Psychiatric Clinics of North America, 14*(1), 21–53, v-vi.

Clausen, J. (2008). Moving minds: Ethical aspects of neural motorprostheses. *Biotechnology Journal, 3*(12), 1493–1501.

Clausen, J. (2009). Man, machine and in between. *Nature, 457*(7233), 1080–1081.

Cole, E. (2007). Direct brain-to-game interface worries scientists. In *Wired Magazine*, September 5. http://www.wired.com/medtech/health/news/2007/09/bci_games?currentPage=all

Demos, J. N. (2005). *Getting started with neurofeedback*. New York: W.W. Norton.

Desney, T. S., & Nijholt, A. (Eds.). (2010). *Brain-computer interfaces: Applying our minds to human-computer interaction*. London: Springer.

Dornhege, G., Millán, J. D. R., Hinterberger, T., McFarland, D. J., & Müller, K. R. (Eds.). (2007). *Toward brain-computer interfacing*. London: MIT Press.

Egner, T., & Gruzelier, J. H. (2003). Ecological validity of neurofeedback: Modulation of slow wave EEG enhances musical performance. *Neuroreport, 14*(9), 1221–1224.

Ellison, K. (2010). Neurofeedback gains popularity and lab attention. In *The New York Times*, October 4. http://www.nytimes.com/2010/10/05/health/05neurofeedback.html?pagewanted=all&_r=0

Evans, J. R., & Abarbanel, A. (Eds.). (1999). *Introduction to quantitative EEG and neurofeedback*. San Diego: Academic Press.

Farah, M. J. (2011). Neuroscience and neuroethics in the 21st century. In J. Illes & B. J. Sahakian (Eds.), *Oxford handbook of neuroethics*. New York: Oxford University Press.

Finger, S. (1994). *Origins of neuroscience: A history of explorations in brain function*. New York: Oxford University Press.

Foks, M. (2005). Neurofeedback training as an educational intervention in a school setting: How the regulation of arousal states can lead to improved attention and behavior in children with special needs. *Educational and Child Psychology, 22*(3), 67–77.

Frank, M. G. (2009). Brain rhythms. In M. D. Binder, H. Nobutaka, & U. Windhorst (Eds.), *Encyclopedia of neuroscience* (pp. 482–483). Berlin: Springer.

Fuchs, T., Birbaumer, N., Lutzenberger, W., Gruzelier, J. H., & Kaiser, J. (2003). Neurofeedback treatment for attention-deficit/hyperactivity disorder in children: A comparison with methylphenidate. *Applied Psychophysiology and Biofeedback, 28*(1), 1–12.

Garfield, C. F., Dorsey, R., Zhu, S., Huskamp, H. A., Conti, R., Dusetzina, S. B., Higashi, A., Perrin, J. M., Kornfield, R., & Alexander, G. C. (2012). Trends in attention deficit hyperactivity disorder ambulatory diagnosis and medical treatment in the United States, 2000–2010. *Academic Pediatrics, 12*, 110–116.

Giordano, J., & DuRousseau, D. (2011). Toward right and good use of brain-machine interfacing neurotechnologies: Ethical issues and implications for guidelines and policy. *Cognitive Technology, 15*(2), 5–10.

Giordano, J., & Olds, J. (2010). On the interfluence of neuroscience, neuroethics and legal and social issues: The need for (N)ELSI. *American Journal of Bioethics-Neuroscience, 1*(4), 12–14.

Gruzelier, J., Egner, T., & Vernon, D. (2006). Validating the efficacy of neurofeedback for optimising performance. *Progress in Brain Research, 159*, 421–431.

Gualtieri, C. T., & Johnson, L. G. (2005). ADHD: Is objective diagnosis possible? *Psychiatry (Edgmont), 2*(11), 44–53.

Hammond, D. C. (2011). What is neurofeedback: An update. *Journal of Neurotherapy, 15*, 305–336.

Heinrich, H., Gevensleben, H., Freisleder, F. J., Moll, G. H., & Rothenberger, A. (2004). Training of slow cortical potentials in attention-deficit/hyperactivity disorder: Evidence for positive behavioral and neurophysiological effects. *Biological Psychiatry, 55*, 772–775.

Hildt, E. (2010). Brain-computer interaction and medical access to the brain: Individual, social and ethical implications. *Studies in Ethics, Law and Technology, 4*, 1–22.

Hirshberg, L. M., Chiu, S., & Frazier, J. A. (2005). Emerging brain-based interventions for children and adolescents: Overview and clinical perspective. *Child and Adolescent Psychiatric Clinics of North America, 14*, 1–19.

Hodgson, K., Hutchinson, A. D., & Denson, L. (2012). Nonpharmacological treatments for 810 ADHD: A meta-analytic review. *Journal of Attention Disorders, 20*(10), 1–8.

Hyman, S. E. (2013). Might stimulant drugs support moral agency in ADHD children?. *Journal of Medical Ethics, 39*(6), 369–370.

Jasper, H. H., Solomon, P., & Bradley, C. (1938). Electroencephalographic analysis of behavior problems in children. *The American Journal of Psychiatry, 95*, 641–658.

Jensen, P. S., Arnold, L. E., Swanson, J. M., et al. (2007). 3-year follow-up of the NIMH MTA study. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*(8), 989–1002.

John, E. R., Prichep, L. S., Ahn, H., Easton, P., Fridman, J., & Kaye, H. (1983). Neurometric evaluation of cognitive dysfunctions and neurological disorders in children. *Progress in Neurobiology, 21*, 239–290.

Kamiya, J. (1968). Conscious control of brain waves. *Psychology Today, 1*, 56–60.

Keizer, A. W., Verment, R. S., & Hommel, B. (2010). Enhancing cognitive control through neurofeedback: A role of gamma-band activity in managing episodic retrieval. *NeuroImage, 49*(4), 3404–3413.

Kennedy, P. (2011). The cyborg in us all. In *The New York Times*, September 14. http://www.nytimes.com/2011/09/18/magazine/the-cyborg-in-us-all.html?pagewanted=all

Kropotov, J. D. (2009). *Quantitative EEG, event-related potentials and neurotherapy*. Amsterdam: Academic.

La Vaque, T. J., & Rossiter, T. (2001). The ethical use of placebo controls in clinical research: The Declaration of Helsinki. *Applied Psychophysiology and Biofeedback, 26*(1), 23–37; discussion 61–65.

Leins, U., Goth, G., Hinterberger, T., Klinger, C., Rumpf, N., & Strehl, U. (2007). Neurofeedback for children with ADHD: A comparison of SCP and theta/beta protocols. *Applied Psychophysiology and Biofeedback, 32*, 73–88.

Lindsley, D. B., & Cutts, K. N. (1940). Electroencephalograms of “constitutionally inferior” and behavior problem children. *Archives of Neurology and Psychiatry, 44*(6), 1199–1212. doi:10.1001/archneurpsyc.1940.02280120046003.

Lofthouse, N., Arnold, L. E., Hersch, S., Hurt, E., & DeBeus, R. (2012a). A review of neurofeedback treatment for pediatric ADHD. *Journal of Attention Disorders, 16*(5), 351–372. doi:10.1177/1087054711427530. Epub 2011 Nov 16.

Lofthouse, N., Arnold, L. E., & Hurt, E. (2012b). Current status of neurofeedback for attention-deficit/hyperactivity disorder. *Current Psychiatry Reports, 14*, 536–542.

Loo, S. K., & Brakley, R. A. (2005). Clinical utility of EEG in attention deficit hyperactivity disorder. *Applied Neuropsychology, 12*(2), 64–76.

Loo, S. K., & Makeig, S. (2012). Clinical utility of EEG in attention-deficit/hyperactivity disorder: A research update. *Neurotherapeutics, 9*, 569–587.

Lubar, J. O., & Lubar, J. F. (1984). Electroencephalographic biofeedback of SMR and beta for treatment of attention deficit disorders in a clinical setting. *Biofeedback and Self-Regulation, 9*, 1–23.

Lubar, J. F., & Shouse, M. N. (1976). EEG and behavioral changes in a hyperkinetic child concurrent with training of the sensorimotor rhythm (SMR): A preliminary report. *Biofeedback and Self-Regulation, 1*(3), 293–306.

Marcus, D. K., & Barry, T. D. (2011). Does attention-eeficit/hyperactivity disorder have a dimensional latent structure? A taxometric analysis. *Journal of Abnormal Psychology, 120*(2), 427–442. doi:10.1037/a0021405.

Molina, B. S. G., Hinshaw, S. P., Swanson, J. M., Arnold, L. E., Vitiello, B., Jensen, P. S., Epstein, J. N., Hoza, B., Hechtman, L., Abikoff, H. B., Elliott, G. R., Greenhill, L. L., Newcorn, J. H.,

Wells, K. C., Wigal, T., Gibbons, R. D., Hur, K., Houck, P. R., & The MTA Cooperative Group. (2009). The MTA at 8 years: Prospective follow-up of children treated for combined type ADHD in a multisite study. *Journal of the American Academy of Child and Adolescent Psychiatry, 48*(5), 484–500. doi:10.1097/CHI.0b013e31819c23d0.

Monastra, V. J. (2005). Electroencephalographic biofeedback (neurotherapy) as a treatment for attention deficit hyperactivity disorder: Rationale and empirical foundation. *Child and Adolescent Psychiatric Clinics of North America, 14*, 55–82.

Monastra, V. J., Lubar, J. F., Linden, M., VanDeusen, P., Green, G., Wing, W., Phillips, A., & Fenger, T. N. (1999). Assessing attention deficit hyperactivity disorder via quantitative electroencephalography: An initial validation study. *Neuropsychology, 13*(3), 424–433.

Monastra, V. J., Lynn, S., Linden, M., Lubar, J. F., Gruzelier, J., & LaVaque, T. J. (2005). Electroencephalographic biofeedback in the treatment of attention-deficit/hyperactivity disorder. *Applied Psychophysiology and Biofeedback, 30*(2), 95–114.

Nijboer F., Allison, B. Z., Dunne, S., Plass-Oude Bos, D., Nijholt, A., & Haselager, P. (2011) A preliminary survey on the perception of marketability of brain-computer interfaces and initial development of a repository of BCI companies. *Proceedings of 5th International Conference on Brain-Computer Interfaces, Technological University Graz*, pp. 344–347.

Nowlis, D. P., & Kamiya, J. (1970). The control of electroencephalographic alpha rhythms through auditory feedback and the associated mental activity. *Psychophysiology, 6*, 476–484. doi:10.1111/j.1469-8986.1970.tb01756.x.

Nuffield Council on Bioethics (2013) Novel neurotechnologies: intervening in the brain. London: Nuffield Council on Bioethics.

Ochs, S. (2004). *A history of nerve function*. Cambridge: Cambridge University Press.

Plischke, H., DuRousseau, D., & Giordano, J. (2011). EEG-based neurofeedback: The promise of neurotechnology and need for neuroethically informed guidelines and policies. *Ethics in Biology, Engineering & Medicine – An International Journal, 2*(3), 221–232.

Polanczyk, G., de Lima, M. S., Horta, B. L., et al. (2007). The worldwide prevalence of ADHD: A systematic review and metaregression analysis. *The American Journal of Psychiatry, 164*, 942–948.

Rothenberger, A., & Rothenberger, L. G. (2012). Updates on treatment of attention-deficit/hyperactivity disorder: Facts, comments, and ethical considerations. *Current Treatment Options in Neurology, 14*(6), 594–607.

Rutger, J. V., Steines, D., Szibbo, D., Kübler, A., Schneider, M.-J., Haselager, P., & Nijboer, F. (2012). Ethical issues in brain–computer interface research, development, and dissemination. *Journal of Neurologic Physical Therapy, 36*(2), 94–99. doi:10.1097/NPT.0b013e31825064cc.

Satterfield, J. H., Cantwell, D. P., Saul, R. E., Lesser, L. I., & Podosin, R. L. (1973). Response to stimulant drug treatment in hyperactive children: Prediction from EEG and neurological findings. *Journal of Autism and Childhood Schizophrenia, 3*(1), 36–48.

Shouse, M. N., & Lubar, J. F. (1979). Operant conditioning of EEG rhythms and Ritalin in the treatment of hyperkinesis. *Biofeedback and Self-Regulation, 4*(4), 299–312.

Singh, I. (2002a). Bad boys, good mothers and the 'miracle' of Ritalin. *Science in Context, 15*(4), 577–603.

Singh, I. (2002b). Biology in context: Social and cultural perspectives on ADHD. *Children and Society, 16*, 360–367.

Singh, I. (2008). Beyond polemics: Science and ethics of ADHD. *Nature Reviews Neuroscience, 9*(12), 957–964.

Singh, I. (2012). VOICES study: Final report. London, UK.

Singh, I. (2013). Not robots: Children's perspectives on authenticity, moral agency and stimulant drug treatments. *Journal of Medical Ethics, 39*(6), 359–366.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Oxford: Appleton-Century.

Skinner, B. F. (1963). Operant behavior. *The American Psychologist, 18*(8), 503–515.

Sterman, M. B., Wyrwicka, W., & Howe, R. (1969). Behavioral and neurophysiological studies of the sensorimotor rhythm in the cat. *Electroencephalography and Clinical Neurophysiology, 27*, 678–679.

Strehl, U., Leins, U., & Goth, G. (2006). Self-regulation of slow cortical potentials: A new treatment for children with attention-deficit/hyperactivity disorder. *Pediatrics, 118*, 1530–1540.

Striefel, S. (2009). Ethics in neurofeedback practice. In T. H. Budzynski, H. K. Budzynski, J. R. Evans, & A. Abarbanel (Eds.), *Introduction to quantitative EEG and neurofeedback*. Amsterdam: Academic.

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., et al. (2001). Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD & ODD symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 168–179.

Swanson, J. M., Wigal, T., & Lakes, K. (2009). DSM-V and the future diagnosis of attention-deficit/hyperactivity disorder. *Current Psychiatry Reports, 11*(5), 399–406.

Tamburrini, G. (2009). Brain to computer communication: Ethical perspectives on interaction models. *Neuroethics, 2*, 137–149.

Thatcher, R. W. (1999). EEG database guided neurotherapy. In J. R. Evans & A. Abarbanel (Eds.), *Introduction to quantitative EEG and neurofeedback*. San Diego: Academic.

Thatcher, R. W., & Lubar, J. F. (2009). History of the scientific standards of QEEG normative databases. In T. H. Budzynski, H. K. Budzynski, J. R. Evans, & A. Abarbanel (Eds.), *Introduction to quantitative EEG and neurofeedback*. Amsterdam: Academic.

van Erp, J. B. F., Lotte, F., & Tangermann, M. (2012). Brain-computer interfaces: Beyond medical applications. *Computer, 45*, 26–34.

van Est, R., van Keulen, I., Geesink, I., Schuijff, M., van den Besselaar, P., Bornmann, L., Leydesdorff, L., Merkx, F., Gurney, T., & van Koten, R. (2010). *Making perfect life: Bio-engineering (in) the 21st century*. Brussels: Rathenau Instituut.

Vernon, D. J. (2005). Can neurofeedback training enhance performance? An evaluation of the evidence with implications for future research. *Applied Psychophysiology and Biofeedback, 30*(4), 347–364.

Weiss, M. D., Baer, S., Allan, A. B., Saran, K., & Schibuk, H. (2011). The screens culture: Impact on ADH. *ADHD Attention Deficit and Hyperactivity Disorders, 3*, 327–334.

World Medical Association. (2000). The declaration of Helsinki. 52nd WMA General Assembly, Edinburgh, Scotland. Retrieved from http://www.wma.net.

WWK. (2008). What we know: AD/HD and neurofeedback: A review of eight studies by the National Resource Center on AD/HD. http://www.help4adhd.org/documents/neurofeedback_8_study_review.pdf.

Yucha, C., & Gilbert, C. (2004). *Evidence based practice in biofeedback & neurofeedback*. Wheat Ridge: AAPB.

Zander, T. O., & Kothe, C. (2011). Towards passive brain–computer interfaces: Applying brain-computer interface technology to human–machine systems in general. *Journal of Neural Engineering, 8(2)*, 025005.

Zuvekas, S. H., & Vitiello, B. (2012). Stimulant medication use in children: A 12-year perspective. *The American Journal of Psychiatry, 169*, 160–166.

Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges

47

Fabrice Jotterand and James Giordano

Contents

Abstract

This chapter focuses on the engagement of real-time functional magnetic resonance imaging-brain-computer interfacing (rtfMRI-BCI) in the treatment of psychopathy and some of the more pertinent ethico-legal and social issues fostered by such use of this neurotechnological approach. To this end, we first provide an overview of the nature of psychopathy. Second, we pose the premise that given the paucity – if not frank absence – of effective psychopharmacological treatment(s) or rehabilitation strategies presently available for psychopathy, it becomes important to examine the present state of neurotechnologies that

F. Jotterand (✉)
Institute for Biomedical Ethics, University of Basel, Basel, Switzerland
e-mail: fabrice.jotterand@unibas.ch

J. Giordano
Neuroethics Studies Program, Pellegrino Center for Clinical Bioethics, Division of Integrative Physiology; Department of Biochemistry and Integrative Program in Neurosciences, Georgetown University Medical Center, Washington, DC, USA

Human Science Center, Ludwig Maximilians Universität, Munich, Germany
e-mail: jg353@georgetown.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_43,

might be used to effect potential benefit in the treatment of this disorder and focus this examination upon the possible utility of rtfMRI-brain-computer interface technology. Third, we present an overview of those tools that are currently used to determine and diagnose psychopathy and discuss their limitations. Finally, we address the major ethical questions and issues arising from the use of this technology to modify behavior in individuals with psychopathic traits.

Introduction

Advances in neuroscience and neurotechnology have increased the capability to assess and manipulate the brain. The use of neurotechnologies to depict, define, and evaluate neural systems' structure and function has expanded the fund of extant knowledge and, as noted by sociologist Bruno Latour (1987), in so doing has prompted additional and perhaps more incising questions about the nature of consciousness, emotions, free will, morality, and constructs of mind and self. Several neurotechnologies are currently employed in clinical, legal, and social contexts; at present, these include (1) neuroimaging, which is being increasingly employed – albeit with considerable criticism – in forensics (e.g., lie detection); (2) neurostimulation technologies, which have been engaged to provide non-pharmacological treatments for a variety of neuropsychiatric disorders (e.g., Parkinsonism, treatment-resistant depression, chronic pain); and (3) those neurotechnologies that are currently in use and being developed to increase cognitive performance (e.g., "cognitive tune-up" [Ong 2008]; "aug-cog" military applications [McBride and Schmorrow 2005; for extensive overview of the use of neurotechnology in military and national security and defense applications, see the thematic issue of *Synesis*: *A Journal of Science*, *Technology*, *Ethics and Policy*, 2010(2)]), to control motor activity, and/or modify behavior (i.e., treatment of psychopathy [Sitaram et al. 2007; Vaadia and Birbaumer 2009]). This chapter focuses on the engagement of real-time functional magnetic resonance imaging-brain-computer interfacing (rtfMRI-BCI) in the treatment of psychopathy and some of the more pertinent ethico-legal and social issues fostered by such use of this neurotechnological approach. To this end, we first provide an overview of the nature of psychopathy. Second, we pose the premise that given the paucity – if not frank absence – of effective psychopharmacological treatment(s) or rehabilitation strategies presently available for psychopathy (Ogloff and Wood 2010; Harris and Rice 2006), it becomes important to examine the present state of neurotechnologies that might be used to effect potential benefit in the treatment of this disorder and focus this examination upon the possible utility of rtfMRI-brain-computer interface technology. Third, we present an overview of those tools that are currently used to determine and diagnose psychopathy and discuss their limitations. Finally, in the spirit of Latour, we address the major ethical questions and issues arising from the use of this technology to modify behavior in individuals with psychopathic traits.

Psychopathy

Psychopathy[1] is a highly investigated personality disorder; yet, to date, an effective treatment for this condition, which is characterized by emotional dysfunction and antisocial behavior, remains lacking (Ogloff and Wood 2010; Harris and Rice 2006). While partially overlapping as a diagnostic construct with antisocial personality disorder (ASPD), the two conditions are, in fact, not identical because ASPD shares a number of comorbid characteristics with other disorders characterized in the *Diagnostic and Statistical Manual* (*DSM*) of the American Psychiatric Association (now in its fifth edition, the *DSM*-5; American Psychiatric Association, www.dsm5.org), and there is a general tendency for overuse of crime-related indicators in ASPD (Widiger 2006). However, while psychopathy is included in the *DSM-5* under the Antisocial/Psychopathic Personality Disorder Type, it is worth noting that individuals with psychopathic traits do not necessarily display criminality. To be sure, psychopaths can manifest antisocial behavior and, by definition, lack emotional response to negative stimuli, yet these traits can and often do occur without prompting criminal activities, as personified by Sam Vaknin in the documentary *I*, *Psychopath* (Lykken 2006; *I*, *Psychopath*, http://topdocumentaryfilms.com/i-psychopath/).

Based on the *Psychopathy Checklist*: *Screening Version PCL*, psychopathy affects approximately 1–2 % of the general population (Neumann and Hare 2008). Symptoms of psychopathy include emotional and interpersonal superficiality, egocentricity and grandiosity, lack of remorse or guilt, lack of empathy, deceitfulness and manipulation, and shallow emotions. The disorder also entails social deviance, manifested in impulsivity, poor behavior controls, need for excitement, lack of responsibility, early behavior problems, and adult antisocial behavior (Hare 1999). According to Kiehl (2006), psychopathy alters cognitive and affective functions in ways that evoke deficiencies in three primary domains: (1) language (i.e., abnormalities in processing abstract or emotional word stimuli), (2) attention and orienting processes (i.e., lack of fear conditioning and unresponsiveness to potential threat(s) of punishment), and (3) affect and emotion (i.e., lack of empathy, guilt or remorse, shallow affect, and irresponsibility and behavioral characteristics such as impulsivity, poor behavioral control, and promiscuity). While only between 15 % and 30 % of the male and female prison population are psychopaths, these individuals have been shown to commit 50 % more crime than non-psychopathic inmates (Sitaram et al. 2007). As well, psychopathy is considered to be a major predictive variable of criminal recidivism. It is estimated that psychopathic offenders are approximately four (4) times more likely to re-offend than non-psychopathic offenders (Hemphill et al. 1998; Fine and Kennett 2004).

Psychopathy is being viewed as paradigmatic pathology for which to use emerging neurotechnologies to define and predict neuropsychiatric disorders of cognition, emotion, and behavior. In this way, the utility and value of neurotechnologies may

[1]This section was adapted with permission from a book chapter: Jotterand et al. (2013)

be (1) in assessing potential predisposition(s) for antisocial and/or criminal acts and (2) used as part of an interventional protocol aimed at preventing such acts. Without doubt, some psychopaths are criminals who need psychiatric treatment. Other individuals have psychopathic traits but have not (as yet) committed criminal acts. Both types of individuals could be a potential threat to society: the former because of their criminality and disposition for recidivism and the latter because of the putative risk(s) fostered by their neuropsychiatric condition. In this light, neurotechnology (e.g., co-registered neuroimaging coupled to neurogenetics and phenotypic markers) could be used to reveal neurobiological traits that may be important to a multicomponential diagnostic protocol, which, if shown to be valid, effective, and reliable, might be ethically and legally justifiable to prompt apprehension and intervention of identified individuals prior to the commission of antisocial, violent acts (Blair 2007, 2008; Birbaumer et al. 2005; Dumit 2003; Deeley et al. 2006; Yang et al. 2009; Blonigen et al. 2003, 2005; Viding et al. 2005; Larsson et al. 2007; Kim-Cohen et al. 2006; Giordano 2011a, b).

As regards intervention, it is important to note that the lack of effective treatment and/or rehabilitative options for psychopathy constitutes a significant challenge for psychiatry, society, and, particularly, for prison administrations. Currently used psychopharmacological interventions are ineffective (Salekin 2002), and subjects who have undergone behavioral therapy fail to exhibit any substantive change in behavior or reduction(s) in recidivism (Abbott 2001; Harris and Rice 2006). Neurotechnologies such as real-time functional magnetic resonance imaging coupled to brain-computer interfaces (rtfMRI-BCI) could provide an alternative approach in the treatment of psychopathy (Sitaram et al. 2007; Vaadia and Birbaumer 2009), yet this is not without limitations, burdens, and potential risks, as we herein seek to illustrate. Although space constraints limit a definitive analysis of rtfMRI, any discussion of rtfMRI-BCI technology, both in general and more specifically for the treatment of psychopathy, must recognize that this approach is in early phases of development and testing, and therefore, further data are required before any conclusive assessment of its viability and ultimate value can be definitively provided. Consequently, this chapter is limited to an overview analysis of the ethical issues arising from the use of rtfMRI-BCI in the determination and diagnosis of psychopathy (and the implications and utility of this neurotechnology in defining subsequent trajectories of intervention), as based on the current information available about both the condition of psychopathy and the state of development and utility of the neurotechnology.

The Possible Utility of Present and Future Neurotechnologies

Functional MRI-BCI constitutes a system that conjoins four major components: the participant, the signal acquisition, the signal analysis, and the signal feedback. Individuals undergoing fMRI-BCI procedures are trained to regulate brain activity by combining "contingent feedback and mental strategies" (Sitaram et al. 2009). This technology-based approach is being developed for the treatment of

psychopathy and other disorders of cognition, emotions, and behavior (e.g., modification of deviant sexual interests in pedophilia [Renaud et al. 2011]). The proposed basis of therapeutic effect(s) is the capacity for individuals to activate and reinforce particular patterns of neural network activity in those brain areas putatively associated with specific types of (favorable or rectified) behaviors (Sitaram et al. 2007; Vaadia and Birbaumer 2009). Various studies using fMRI-BCI technology have provided evidence of behavioral changes in neural network activity in brain areas associated with the perception of pain (e.g., the rostral anterior cingulate cortex [deCharms et al. 2005]), language processing (e.g., the right inferior frontal gyrus [Rota et al. 2008]), and response to certain emotional stimuli (e.g., the anterior insula [Caria et al. 2006, 2007; Sitaram et al. 2009]). A 2005 study by Birbaumer et al. revealed that criminal psychopaths showed a lack of metabolic activity in brain areas associated with mediating fear that are usually activated during aversive classical conditioning (i.e., fear of punishment in which fear conditioning is essential for behavioral development; Birbaumer et al. 2005). It has been hypothesized that fMRI-BCI technology could be utilized to condition psychopaths to activate brain regions involved in fear-evoking situations and, in this way, engage a form of neurofeedback that may be effective in reducing psychopathic traits (Birbaumer and Cohen 2007; Sitaram et al. 2007). This hypothesis is fortified by a 2007 study demonstrating that criminal psychopaths can be conditioned to self-regulate emotions through feedback-induced alteration of the blood oxygen level (BOLD) signal (and by implication, neural network activity) in the anterior insular cortex (Sitaram et al. 2007, 2009).

These findings – while limited and nascent – are suggestive, if not promising, and current interest in this area mandates further analysis of the potential use of rtfMRI-BCI, and other neurotechnologies, to alter the human predicament (of disease, illness, suffering, and sadness), the human condition, and perhaps human beings as well (Jotterand 2008a, b, 2010; Giordano 2010, 2012a, b, d). The use of rtfMRI-BCI for regulating "moral emotions" (Charland 2009) suggests the possibility of employing neurotechnologies to modify human thought and action. Studies have shown that human emotions and behavior may, in fact, be manipulated through a variety of neuroscientific approaches, including systemic and intraneural administration of a range of drugs and neuromodulatory agents (see Harmer et al. 2003; Kosfeld et al. 2005 for specific studies) and the use of neurotechnological devices (such as transcranial magnetic stimulation (TMS) deep brain stimulation (DBS) and vagal nerve stimulation; Horstman 2010; Ong 2008; Higgins and George 2009). Given these findings, it has been suggested that similar approaches could potentially be used to influence moral behavior (Blair 2007). Thus, we could envision trajectories toward neuro-engineering "morally better people" and/or reforming "morally deficient people" (e.g., psychopaths or incarcerated criminals; see Sitaram et al. 2007). Obviously, important steps toward the development and use of any such approaches would entail (1) reliable assessment, diagnosis, and/or prediction of aberrant cognition, emotions, and behaviors and (2) the use of neurotechnology to accurately depict the neural substrates and mechanisms subserving these cognitions and behaviors.

Tools for Diagnosing Psychopathy and Their Limitations

At present, three diagnostic tools are available for the determination of psychopathic traits or clinical psychopathy: (1) the Psychopathy Checklist–Revised (PCL-R), (2) neuroimaging techniques, and (3) neurogenetics. Robert Hare developed the Psychopathy Checklist – Revised (PCL-R) (Hare 1991) that empirically established the validity of a rating scale for psychopathy. Hare's empirical studies segregate four symptom clusters representative of psychopathy: *Interpersonal*: manifested by presentation of glibness/superficial charm, grandiose sense of self-worth, pathological lying, and cunning/manipulative traits; *Affective*: evidenced in lack of remorse or guilt, shallow affect, callousness, lack of empathy, and failure to accept responsibility; *Lifestyle*: a need for stimulation, proneness to boredom, parasitic lifestyle, lack of realistic long-term goals, impulsivity, and general irresponsibility; and *Antisocial*: effected by poor behavioral controls, early behavioral problems, juvenile delinquency, revocation of conditional release, and criminal versatility.

Neuroimaging technologies may provide additional means to determine neuroanatomical variations in brain structure and function that may be reflective – or predictive – of psychopathy (Harenski et al. 2010). Structural magnetic resonance imaging (sMRI) measures brain volume as based upon the paramagnetic signal expressed by relative densities of gray and white matter and provides fine spatial resolution and replicable data sets. However, a limitation of sMRI is that the image is static, which, while (rather precisely) depicting brain structure, does not illustrate any functional parameters. Functional magnetic resonance imaging (fMRI) measures blood flow in specific brain regions through depiction (and statistical parametric measurement, conversion, and analysis) of the differential paramagnetic signal produced by oxygenated and non-oxygenated hemoglobin (i.e., blood-oxygen level-dependent response, BOLD) as a function of variations in regional cerebral oxygen demand evoked by activity within particular neural networks in various brain regions. While this approach offers excellent spatial resolution, it affords only moderate (if not poor) temporal resolution in light of the time frame of neural vs. cerebrovascular events and consequences (NB: this is reflected in the caveat regarding fMRI-based assessment that is colloquially expressed as "fMRI measures *vein* activity, rather than *brain* activity"). In contrast, electroencephalographic recording/event-related potentials (EEG/ERP) assess the electrical activity of the brain through extracranial electrodes and provide superior temporal resolution, albeit with poor spatial resolution (a detailed examination of EEG-based techniques is beyond the scope of this chapter; for overview, see Niedermeyer and da Silva 2004; Huizenga et al. 2001; Hämäläinen et al. 1993; and for discussion of use and utility in forensic psychiatry, see Puranik et al. 2009).

Despite relative constraints and limitations, these techniques have proven useful in showing aberrant patterns of activity in neural networks, in particular brain regions of individuals with psychopathic tendencies (Blair 2008). Recent neuroimaging studies support that the abnormal psychophysiological and cognitive characteristics of psychopathy have definable neurobiological correlates (Pridmore et al. 2005; Raine and Yang 2006). Such investigations suggest that damage to the orbitofrontal

cortex may be associated with cognitive impairments (Kiehl 2006) and insult to the anterior cingulate can lead to emotional blunting, hostility, irresponsibility, and disagreeableness. Furthermore, abnormalities in the structure and/or function of specific regions of the amygdala, and afferent and efferent networks involving the amygdala, may subserve inappropriate responses to affective stimuli such as anger and/or fear expressed by others, which is often a cardinal trait of psychopathic individuals (Mesulam 2000; Aggleton 1992; Anderson et al. 2000; cited in Kiehl 2006). Taken together, results such as these support that neuroimaging may be of value in assessing and detecting neuroanatomic and neurophysiological variables that are diagnostic and perhaps predictive of aberrant cognitions, emotions, and behaviors that are representative of psychopathy.

Studies have also suggested a possible genetic contribution to psychopathy (Harenski et al. 2010), although the nature of genetic-environmental interactions and effects are not well defined (see, e.g., Ridley 2003, for review). Twin studies using the Psychopathic Personality Inventory revealed moderate-to-high heritability of psychopathic traits (Blonigen et al. 2003; see also Blonigen et al. 2005). A twin study by Viding and colleagues (2005) investigating the heritability of callous-unemotional traits and antisocial behavior suggested an influence of genetic factors in behavioral characteristics, and genetic studies found association between functional polymorphisms in the monoamine oxidase A (MAO-A) gene and conduct disorder (Harenski et al. 2010; see also analysis by Kim-Cohen et al. 2006). Other studies have examined correlation of MAO-A levels and traits that are representative of and/or prevalent in psychopathy (such as aggression and impulsivity; Buckholtz and Meyer-Lindenberg 2008), and a neuroimaging-genetic study found that males with low levels of MAO-A gene and enzyme had lower amygdalar volume and increased ventromedial prefrontal volume as compared to males with higher levels of MAO-A gene and enzyme (Meyer-Lindenberg et al. 2006). These studies infer that particular genetic variations may be involved in, and contributory to anatomical features that are observed in psychopathy and that these genetic and anatomical features may be validly and reliably assessable using current (and planned) neuroscientific and neurotechnological means. However, caution is warranted when proposing any form of definitive genetic and anatomic reductionism, i.e., establishing a wrong correlation between genes and behavior. Given the polymorphic and/or pleiotropic (multiple phenotypic) nature of neuropsychiatric disorders, the presence of specific genes is not sufficient to establish the primary cause of behavioral and/or psychological traits, as multiple genetic, anatomical, and environmental factors and interactions are critical (both at key developmental periods and to varying extent throughout the life span) to the spectral expression of a condition or disorder, inclusive of psychopathy (Giordano and Wurzman 2008; FitzGerald and Wurzman 2010; Wurzman and Giordano 2012).

These caveats are important in light of (1) an increasing trend toward employing neuroscientific and neurotechnological approaches such as neuroimaging and neurogenetics to define, assess, and determine a variety of cognitive, emotional, and behavioral characteristics; (2) existing strengths and limitations of these approaches; and (3) a call to use such approaches in ways that can prevent potentially

violent manifestations of particular psychiatric disorders, such as psychopathy. Thus, we urge the need to realistically and clearly elucidate (1) if and how such approaches might be used to validly and reliably detect and perhaps predict psychopathic traits and (2) what criteria must be developed, addressed, and satisfied in order to ethically and legally substantiate using neurotechnologies in these ways. These requirements compel the type of broad analysis of the neuroscience of psychopathy and the neuroscientific and neurotechnological approaches that could be employed to assess and intervene against the socially destructive traits of this disorder that we cannot provide given the space allocated. To wit, in what follows, we will focus our examination upon the ethical challenges arising from the proposed use of rtfMRI-BCI to assess – and shape intervention against – psychopathy.

Potential Ethical Concerns

The ability to assess, access, and manipulate behavior through rtfMRI-BCI has been advanced by a limited number of studies that have demonstrated positive behavioral changes (Caria et al. 2012). Thus, while these studies are restricted in scope, the implications drawn from their results are that this neurotechnology represents a promising, complementary, and perhaps even alternative approach to the use of existing, rather ineffective, interventions. Beyond these positive developments, further, more detailed and longitudinal studies are required to ascertain both the effectiveness and viability of this neurotechnological approach and to consider important issues that are related to current neuroscientific research on psychopathy, the technical and methodological challenges posed when attempting to utilize neuroscientific techniques and tools in assessing and treating this disorder, and the ethico-legal and social implications generated by these approaches. As with other new and cutting-edge technologies, there is considerable risk of overstating the potential applications and effectiveness of rtfMRI-BCI, and this incurs a number of ethical issues and problems. Questions such as honesty and veracity about the actual capabilities (and limitations), and the intended goals of employing rtfMRI-BCI, require careful analysis and articulation, so as to avoid misrepresenting what the technology can truly achieve and what various user groups (e.g., within medical institutions, prisons etc.) can offer to those treated and to society at large. For example, particular methodological issues related to the assessment of behavioral effects while using rtfMRI-BCI technology pose a twofold challenge, namely, (1) what methods can and should be used to validly correlate the activity and regulation of neural networks to specific cognitive, emotional, and motor effects and (2) can – and/or how can – subjects be conditioned to regulate BOLD responses to affect patterns of neural network activity and particular cognitions, emotions, and behaviors (Sitaram et al. 2009)? These questions, while fundamental to the value of rtfMRI-BCI-based approaches in assessing and treating psychopathy, are not insurmountable. Instead, we posit that issues such as these provide sentinel direction(s) for continuing research toward more finely grained elucidation of the substrates and

mechanisms of psychopathy and how this technology – and others – might be (best) used for neural network activation and behavioral control.

Still, there are additional challenges that raise ethical and social questions about such use of rtfMRI-BCI: Neuroimaging studies have been performed on relatively few psychopaths because of the difficulty in recruiting individuals that "truly" present with this disorder (i.e., exceed a threshold of 30 or higher on the PCL-R) (Harenski et al. 2010). As well, psychopaths often manifest comorbidities such as substance abuse and dependence, which can alter function of certain brain regions (e.g., the orbitofrontal cortex) that have been implicated in neuroanatomically based diagnoses of psychopathy (Harenski et al. 2010). Furthermore, developmental and longitudinal neuroimaging studies that would be important, if not necessary, to determine possible neural variation(s) and abnormalities in psychopathy are lacking (Harenski et al. 2010). To this point, it remains unclear whether neurobiological variations are the *cause* of psychopathy, whether they are an *effect*, or whether they are simply epiphenomenal. Central to this issue is the question of whether and to what extent genetic influence and/or the development of – or insult to – neuroanatomical structures and functions contribute to, and can be assessed as definitive and predictive of, psychopathy. Blair has asserted that there is always risk that "...a brain scan diagnosis of psychopathy legitimizes the preventive incarceration of a 'high-risk' individual, and in which a static neuro-structural deficit may lead to a therapeutically nihilistic approach to such an individual on the grounds that he is beyond rehabilitation" (Blair 2003, p. 564). Considering the lack of effective treatment for psychopathy, and current limitations of neurogenetics and neuroimaging, Blair's statement well-defines potential pitfalls of employing currently available neurogenetic and neuroanatomical methods – either alone or in tandem. Yet, recent events (e.g., Columbine, Phoenix, and Connecticut shootings) have prompted renewed public interest in – and calls for – the use of neuroscientific, neurotechnological, and psychiatric measures to better define, if not predict and prevent, violent psychopathic behaviors. The key questions then are whether and how neurotechnologies such as rtfMRI-BCI can – and should – be used for social control and to protect the public and what measures of analysis and scrutiny are necessary and sufficient to enable (technical and ethico-legally) prudent interventions. In other words, as neurotechnologies are developed and advanced, and the scope of their application increases, various stakeholders will need to determine (1) whether society has a moral obligation and legal duty to protect its citizenry through the development and use of neurotechnologies to "fix the defective brain" of psychopaths and (2) the source and validity of metrics that undergird and uphold social justification for (a) the use of predictive neurotechnologies, (b) the use of preventive measures prior to the commission of frank social violence, and (c) the enforced treatment of criminal psychopaths and/or enforced diagnosis (and uses of preventive interventions) in noncriminal psychopaths. The absence of effective psychopharmacological treatment and/or rehabilitation strategies also raises issues of how to determine and establish the ethical parameters (and boundaries) needed to (1) enforce therapy as condition for parole and decreased sentencing (of criminal psychopaths), (2) engage a sanctioned program of neuroenhancement (i.e., to

"neuro-engineer" the abolishment of psychopathic traits in those who could pose a threat to society), or (3) institute preventive measures so as to detect and avert socially aggressive and violent behavior in previously undiagnosed psychopaths. This in turn will require an assessment of current governmental regulations and the norms and values required (and used) to secure the safe utilization of neuroscientific and neurotechnologically based sociomedical interventions.

Personal Identity

The technical, methodological, and social questions and ethical implications raised in the previous section demand comprehensive investigation so as to avoid the misuse or abuse of neurotechnology. The focus on psychopathy increases the level of complexity because of the ongoing debate about the nature of the condition, i.e., how and to what degree it affects individuals who may or may not engage in criminal acts, and the moral and legal status of criminal psychopaths. This latter point necessitates questioning whether criminal psychopaths are responsible individuals, albeit diminished in their moral sensitivity due to neuroanatomical abnormalities and genetic determinants, thereby deserving criminal sentencing, or whether they are what Fine and Kennett refer to as "blameless offenders" (Fine and Kennett 2004) who instead should be required to undergo treatment in light of the "disease" or "disorder" of abnormal brain structure and/or function (see Jotterand et al. 2013, in press) for further analysis of this point).

The formulation of meaningful answers to these questions is dependent to a large part upon continuing neuroscientific research that is dedicated to further an understanding of psychopathy. An unresolved ethical question inherent to any such research – and its outcomes – relates to the implications of using neurotechnologies in diagnosis, establishment of medical, social, and legal ontologies, and the impact of such information and constructs on the concept (and value) of personal identity. The psychiatric and neuroethics literature has already become focused on questions of enforced therapy, neuroenhancement, and preventive measures (Matravers 2010; Glannon 2011; Gordijn and Chadwick 2008), and in many ways, this discourse provides the groundwork for consideration of rtfMRI-BCI as discussed here. For example, the issue of how TMS and DBS could affect personal identity has been addressed elsewhere (Jotterand and Giordano 2011), but we assert that it is also worthwhile elaborating upon those deliberations to date and the specific aspects that are most relevant to a discussion of the use of rtfMRI-BCI to assess and treat psychopathy.

Such questions confront contemporary and long-standing notions of self and embodiment and the relation of mind to brain. To approach these questions, we hold that the concept of embodiment is essential to an understanding of how individuals construe their sense of personal identity and/or notion of self. Individuals are shaped by experiences that are interpreted through the medium of the body, which is situated in space and time (Merleau-Ponty 1945). This means that personal identity comprises the aspects of human development and experience that are both biological and psychosocial (Jotterand and Giordano 2011).

It is one thing to establish the importance of embodiment for a definition of personal identity; however, the relationship between the perceptual body (and its instrument for the interpretation of perception, the brain) and the entity referred to as the "mind" still remains unresolved. This issue can be approached from philosophical, neurophysiological, and neuropsychological perspectives and each provide particular insights that putatively describe aspects of the relationship of brain to mind (Churchland 1990; Damasio 1999; Prinz 2005; Heidegger 1962; Jaspers 1949; and Merleau-Ponty 1945). Obviously, we will not settle the question in this chapter; however, we do wish to stress that human experience in the world is defined by bodily sensation and phenomenal cognition (not in a dualistic, Cartesian sense, but rather as a complementary dyad of somatic and psychological reality) (Giordano 2008a; Shook and Giordano 2013), which then prompts important questions about brain interventions that affect mental states. These interventions have the potential to disrupt the integrity of the self, which could be morally questionable in the therapeutic application of neurotechnology if such interventions undermine individuals' autonomy (self-determination) and privacy (protection of health information including the predictive and personal nature of neurobiological information) (Jotterand and Giordano 2011). One problem is that the use of neuroscientific and neurotechnological approaches to "predict" the potential occurrence of a disorder such as psychopathy (that fosters considerable socio-legal concern and consternation) could result in bias(es). As well, using the nature and manifestations of the disorder (i.e., as a potential threat to society) in order to justify interventions without proper consideration of the (integrity of the) personal identity of psychopaths or individuals with psychopathic traits, is ethico-legally problematic.

Assuming, hypothetically, that rtfMRI-BCI technology were to become an established option for treating psychopathy, we predict that social expectations to use such technology would accordingly increase, particularly in light of the recent events of terrorist acts and shootings, and the previously mentioned social pleas for neuroscientific and neurotechnological means of intervention and prevention. Additionally, rtfMRI-BCI might be used to manipulate the brain∼minds of people with psychopathic traits, without the safeguards of appropriate consent. This concern, however, can be quickly dismissed for two ethical and practical reasons. First, the principle of informed consent protects competent subjects (including prison inmates) from frank manipulation and coercion. But let us not be naïve; of course, the potential for coercing inmates may in fact be greater because of their status and background. That said, the principle of equivalence as applied in prison medicine provides a fair measure of surety as it entails (at least in the ideal) that prison populations should receive the same medical care as free individuals, inclusive of the ethical obligation to obtain informed consent. Second, rtfMRI-BCI requires the cooperation of subjects (who want to undergo treatment in order to change their behavior). As previously stated, rtfMRI-BCI uses a type of neurofeedback that demands active participation of subjects in the conditioning process. Coercion and manipulation undermine the quality of therapeutic interventions. Therefore, while ascription to mandatory therapy could offer incentives to criminal psychopaths as conditions for parole, early release or shorter sentencing – if studies show the safety and effectiveness of

rtfMRI-BCI-based interventions – it is paramount to have the full consent of these subjects to achieve the highest degree of effectiveness in the therapy.

Perhaps the real ethical issue is not mandatory treatment of convicted psychopaths, but instead the longitudinal use of (this) neurotechnology. As with certain types of genetic testing (Fulda and Lykens 2006; Hodge 2004), pediatric brain scanning might become mandatory and, depending upon a child's subsequent behavioral development, could be used as a metric to establish thresholds and criteria for pediatric rtfMRI-BCI treatment – as a type of "preventive therapy" or, perhaps more or less accurately, a form of socio-moral neuroenhancement (vide supra). Such treatment would require (substituted) consent of parents but also the assent of a child or young (i.e., less than "legal age") adolescent. Intervention would be presented as a therapeutic good that would enhance the quality of life of the subject, and as a social good that seeks to preserve the safety of the public at large. But we must ask, just how informed could we expect subjects to be, given the nascence and novelty of the technology and the changing epistemic capital of neuroscience? And given these same contingencies, and their influence upon the knowledge base of clinicians, legal practitioners, and/or administrative and political personnel responsible for establishing protocols and precedents, it is important to assess what information is necessary and sufficient to provide to patients about new scientific techniques and technologies and if and how the use (or nonuse) of such approaches is important to clinical equipoise (Giordano 2011a, b). Moreover, given the reliance upon computational informatics (in the comparative evaluation and normative indexing of neuroimaging-based diagnostics; see Giordano 2012d, for overview), we must question how private any information relevant to neurotechnologically based diagnoses and interventions for psychopathy will actually remain (e.g., in the face of possible use of these data for commercial and actuarial purposes) and what burdens, risks, and possible harms such (unintended or in some cases intentional) diffusion of information will incur (i.e., to individuals, communities, and society; Giordano 2011a, b).

The issue of preventive measures raises further questions of whether and in what ways such treatment represents a violation of individual autonomy. In principle, an intervention rendered as enforcement of the law is justifiable only when individuals engage in criminal or socially injurious behaviors (Morse 2010). While common standard of law asserts that absent a determination of guilt, there is no mechanism to (legally) mandate preemptive "treatment," there is a building sentiment – and increasing call (from certain factions of the public, professional and political communities) – for the use of neuroscience and neurotechnology to be employed to "protect the public" that could be used to support arguments for such predictive and preventive approaches. Of course, this stance is contentious, and ongoing debate is focused upon the utility of the spectrum-disorder concept in making predictive inference about future trajectories of bio-psychosocial effects (see, e.g., Giordano and Wurzman 2008; Wurzman and Giordano 2012, for overviews) and representation of the validity, relative value(s), burdens, risks, and harms of this approach in light of outcomes achieved to date. But here we advocate extreme caution; it will be increasingly important to weigh the merit of each type and

application of neuroscientific techniques and technology relative to assessing and "treating" particular conditions, states, and traits in specific individuals. Without doubt, a case-by-case approach should be utilized, but further, it will be important to examine the viability of extant ethical and legal criteria (such as Frye and Daubert standards) and recognize the tendency to use extant technology – and ethico-legal justifications – in ways that are inapt (Giordano 2010, 2012b). It may be that the presence of such neuroscientific techniques and technologies within the legal, economic, and political armamentarium warrants a deeper and more intricate examination, evaluation, and possible reformulation of certain ethico-legal constructs, guidelines, and pre-/proscriptions, and this will constitute something of a sea change (Benedikter and Giordano 2011; Giordano 2011a, b, 2012a–d; Giordano and Benedikter 2010, 2012; Shook and Giordano 2013, in press).

Conclusion

Psychopathy not only affects the individual afflicted but can – and frequently does and will – impact the lives of others: the victims of psychopaths' socially aggressive and violent behaviors as well as their families, friends, and, in many ways, society at large. The challenges posed by the condition, diagnosis, treatment, and prevention of psychopathy to existing constructs of moral agency and criminal responsibility are fortified by the availability and putative viability of ever newer techniques and technologies of neuroscience. The heretofore absence of effective treatments has afforded strong incentives for the development, testing, and implementation of neurotechnologies, such as rtfMRI-BCI, for the assessment and treatment of psychopathy. It may be, as we have noted both in this chapter and elsewhere, that neuroscience and neurotechnology offer real promise and potential to alleviate the individual and social burdens incurred by disease, injury, illness, environments, and age; yet such promise is accompanied by potential problems and perils. In light of this, we have advocated a preparatory stance that is well informed of the practical, ethico-legal, and social benefits, limitations, risks, and implications that neuroscientific and technological progress may offer and incur (Benedikter and Giordano 2011; Giordano 2008b, 2009, 2010, 2012a, b, c, d; Giordano and Benedikter 2011, 2012; Giordano and DuRousseau 2011; Plischke et al. 2012). But a preparatory stance does not necessarily prevent the advancement and use of neuroscientific techniques and technology (Benedikter and Giordano 2012; Giordano et al. 2010); thus, we posit that it will be paramount to determine the ethical and legal criteria with which to evaluate if and what methods and protocols of neurotechnological assessments can be employed to initiate social interventions to mitigate and/or prevent psychopathy and its predisposing traits. This in turn will require establishing thresholds for both diagnosis and any subsequent interventions and, beyond techno-scientific and methodological considerations, will mandate ongoing address, analysis, and revision of existing ethical and legal guidelines that can be engaged to direct and govern the medical, legal, and political use of neuroscience and neurotechnology. We argue that this is both a challenge and opportunity for the field

and practice of neuroethics, not as a stand-alone discipline, but as a critical constituent of a larger infrastructure of medicine, law, and social administration that is responsive and dedicated to individual and public good (Giordano 2010, 2011a; 2011b; Giordano and Olds 2010). It is our hope that this project of neuroethics, legal, and social engagement remains a strongly positive work in progress.

Acknowledgements The authors gratefully acknowledge the assistance of Daniel Howlader in the preparation of this manuscript. This work was funded in part by the J. W. Fulbright Foundation (JG), Clark Fellowship in Neuroscience and Ethics (JG), William H. and Ruth Crane Schaefer Endowment (JG), and ongoing support from the Neuroethics Studies Program of the Pellegrino Center for Clinical Bioethics, Division of Integrative Physiology/Department of Biochemistry, and Graduate Liberal Studies Program, Georgetown University, Washington DC, USA (JG).

Cross-References

- ▶ Brain Research on Morality and Cognition
- ▶ Determinism and Its Relevance to the Free-Will Question
- ▶ Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine
- ▶ Drug Addiction and Criminal Responsibility
- ▶ Ethical Implications of Brain–Computer Interfacing
- ▶ Ethical Implications of Brain Stimulation
- ▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
- ▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
- ▶ Free Will, Agency, and the Cultural, Reflexive Brain
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
- ▶ Moral Cognition: Introduction
- ▶ Neuroenhancement
- ▶ Neuroethics and Identity
- ▶ Prediction of Antisocial Behavior
- ▶ Reflections on Neuroenhancement
- ▶ Research in Neuroenhancement
- ▶ The Morality of Moral Neuroenhancement

References

Abbott, A. (2001). Into the mind of a killer. *Nature, 410*, 296–298.

Aggleton, J. P. (1992). The functional effects of amygdala lesions in humans: A comparison with findings in monkeys. In J. P. Aggelton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and depression*. New York: Wiley-Liss.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Anderson, A. K., Spencer, D. D., Fulbright, R. K., & Phelps, E. A. (2000). Contribution of the anteromedial temporal lobes to the evaluation of facial emotion. *Neuropsychology, 144*, 526–536.

Benedikter, R., & Giordano, J. (2011). The outer and inner transformation of the global sphere through technology: The state of two fields in transition. *New Global Studies, 5*(2), 1–17.

Benedikter, R., & Giordano, J. (2012). Neurotechnology: New frontiers for European policy. *Pan European Networks: Science & Technology, 3*, 204–207.

Birbaumer, N., & Cohen, L. G. (2007). Brain-computer interfaces: Communication and restoration of movement in paralysis. *The Journal of Physiology, 579*, 621–636.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., & Flor, H. (2005). Deficient fear conditioning in psychopathy: A functional magnetic resonance imaging study. *Archives of General Psychiatry, 62*, 799–805.

Blair, R. J. R. (2003). Neuroimaging psychopathy: Lessons from Lombroso. *The British Journal of Psychiatry, 182*, 5–7.

Blair, R. J. R. (2007). The amygdale and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences, 11*, 387–392.

Blair, R. J. R. (2008). The cognitive neuroscience of psychopathy and implications for judgments of responsibility. *Neuroethics, 1*, 149–157.

Blair, R. J. R. (2010). Neuroimaging of psychopathy and antisocial behavior: A targeted review. *Current Psychiatry Reports, 12*, 76–82.

Blonigen, D. M., Carlson, S. R., Krueger, R. F., & Patrick, C. J. (2003). A twin study of self-reported psychopathic personality traits. *Personality and Individual Differences, 35*, 179–197.

Blonigen, D. M., Hicks, B. M., Krueger, R. F., Patrick, C. J., & Iacono, W. G. (2005). Psychopathic personality traits: Heritability and genetic overlap with internalizing and externalizing psychopathology. *Psychological Medicine, 35*, 637–648.

Buckholtz, J., & Meyer-Lindenberg, A. (2008). MAOA and the neurogenetic architecture of human aggression. *Trends in Neuroscience, 31*, 120–129.

Caria, A., et al. (2006). Can we learn to increase our emotional involvement? Real-time fMRI of anterior cingulate cortex during emotional processing. *Human Brain Mapping.*

Caria, A., et al. (2007). Regulation of anterior insular cortex activity using real-time fMRI. *NeuroImage, 35*, 1238–1246.

Caria, A., Sitaram, R., & Birbaumer, N. (2012). Real-time fMRI: A tool for local brain regulation. *The Neuroscientist, 18*(5), 487–501.

Charland, L. C. (2009). Technological reason and the regulation of emotion. In J. Phillips (Ed.), *Philosophical perspectives on technology and psychiatry*. Oxford: Oxford University Press.

Churchland, P. S. (1990). *Neurophilosophy: Toward a unified science of the mind/brain*. Cambridge, MA: MIT Press.

Damasio, A. (1999). *The feeling of what happens – body and emotion in the making of consciousness*. New York: Harcourt.

deCharms, R. C., Meada, F., Glover, G. H., Ludlow, D., Pauly, J. M., Soneji, D., et al. (2005). Control over brain activation and pain learned by using real-time functional MRI. *Proc Natl Acad Sci USA, 102*, 18626–18631.

Deeley, Q., Daly, E., Surguladze, S., Tunstall, N., Mezey, G., Beer, D., et al. (2006). Facial emotion processing in criminal psychopathy. Preliminary functional magnetic resonance imaging study. *The British Journal of Psychiatry, 189*, 533–539.

Dumit, J. (2003). *Picturing personhood: Brain scans and biomedical identity*. Princeton: Princeton University Press.

Fine, C., & Kennett, J. (2004). Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment. *International Journal of Law and Psychiatry, 27*, 425–443.

FitzGerald, K., & Wurzman, R. (2010). Neurogenetics and ethics. In J. Giordano & B. Gordijn (Eds.), *Scientific and philosophical perspectives in neuroethics*. Cambridge: Cambridge University Press.

Fulda, K. G., & Lykens, K. (2006). Ethical issues in predictive genetic testing: A public health perspective. *Journal of Medical Ethics, 32*, 143–147.

Giordano, J. (2008a). Complementarity, brain-mind, and pain. *Forschende Komplementärmedizin, 15*, 2–6.

Giordano, J. (2008b). Technology in pain medicine: Research, practice, and the influence of the market. *Prac Pain Management, 8*, 56–59.

Giordano, J. (2009). The intersection of ethics, education and policy. *Prac Pain Management, 9*, 63–67.

Giordano, J. (2010). The mechanistic paradox: The relationship of science, technology, ethics and policy. *Synesis, 1*, G1–G4.

Giordano, J. (2011a). Neuroethics: Traditions, tasks and values. *Human Prospect, 1*, 2–8.

Giordano, J. (2011b). Neuroethics- two interacting traditions as a viable meta-ethics? *American Journal of Bioethics – Neuroscience, 3*, 23–25.

Giordano, J. (2012a). Integrative convergence in neuroscience: Trajectories, problems and the need for a progressive neurobioethics. In A. Vaseashta, E. Braman, & P. Sussman (Eds.), *Technological innovation in sensing and detecting chemical, biological, radiological, nuclear threats and ecological terrorism (NATO Science for Peace and Security Series)*. New York: Springer.

Giordano, J. (2012b). Neuroimaging in psychiatry: Approaching the puzzle as a piece of the bigger picture(s). *American Journal of Bioethics – Neuroscience, 3*, 54–56.

Giordano, J. (Ed.). (2012c). *Neurotechnology: Premises, potential and problems*. Boca Raton, FL: CRC Press.

Giordano, J. (2012d). Unpacking neuroscience and neurotechnology – instructions not included: neuroethics required. *Neuroethics, 4*.

Giordano, J., & Benedikter, R. (2011). The shifting architectonics of pain medicine: Toward ethical re-alignment of scientific, medical and market values for the emerging global community – groundwork for policy. *Pain Medicine, 12*, 406–414.

Giordano, J., & Benedikter, R. (2012). An early – and necessary – flight of the Owl of Minerva: Neuroscience, neurotechnology, human socio-cultural boundaries, and the importance of neuroethics. *Journal of Evolution and Technology, 22*, 14–25.

Giordano, J., & DuRousseau, D. (2011). Toward right and good use of brain-machine interfacing neurotechnologies: Ethical issues and implications for guidelines and policy. *Cognitive Technology, 15*, 5–10.

Giordano, J., & Gordijn, B. (Eds.). (2010). *Scientific and philosophical perspectives in neuroethics*. Cambridge: Cambridge University Press.

Giordano, J., & Olds, J. (2010). On the interfluence of neuroscience, neuroethics, legal and social issues: The Need for (N)ELSI. *Am J Bioethics Neurosci, 2*(2), 13–15.

Giordano, J., & Wurzman, R. (2008). Neurological disease and depression: The possibility and plausibility of putative neuropsychiatric spectrum disorders. *Depression: Mind and Body, 4*, 2–5.

Giordano, J., Forsythe, C., & Olds, J. (2010). Neuroscience, neurotechnology, and National security: The need for preparedness and an ethics of responsible action. *American Journal of Bioethics – Neuroscience, 1*, 35–36.

Glannon, W. (2011). *Brain, body, and mind*. Oxford: Oxford University Press.

Gordijn, B., & Chadwick, R. (Eds.). (2008). *Medical enhancement and posthumanity*. Dordrecht: Springer.

Hämäläinen, M. R., Hari, R., Ilmoniemi, J. K., & Lounasmaa, O. V. (1993). Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics, 65*, 413–497.

Hare, R. D. (1991). *Manual for the hare psychopathy checklist-revised*. Toronto: Multi-Health Systems.

Hare, R. D. (1999). *Without conscience: The disturbing world of the psychopath among us*. New York, NY: Guilford Press.

Harenski, C. L., Hare, R. D., & Kiehl, K. A. (2010). *Neuroimaging, genetics, and psychopathy: Implications for the legal system. In L. Malatesti & J McMillan (Ed.) Responsibility and psychopathy: Interfacing law, psychiatry and philosophy*. London: Oxford University Press.

Harmer, C. J., Bhagwagar, Z., Perrett, D. I., Völlm, B. A., Cowen, P. J., & Goodwin, G. M. (2003). Acute SSRI administration affects the processing of social cues in healthy volunteers. *Neuropsychopharmacology, 28*, 148–152.

Harris, G. T., & Rice, M. E. (2006). Treatment of psychopathy: A review of empirical findings. In C. J. Patrick (Ed.), *Handbook of psychopathy*. New York: The Guilford Press.

Heidegger, M. (1962). *Being and time*. (trans: Macquarrie, J., & Robinson, E.). New York: Harper & Row.

Hemphill, J. F., Hare, R. D., & Wong, S. (1998). Psychopathy and recidivism: A review. *Legal and Criminological Psychology, 3*, 139–170.

Higgins, E. S., & George, M. S. (2009). *Brain stimulation therapies for clinicians*. Arlington, VA: American Psychiatric Publishing.

Hodge, J. G. (2004). Ethical issues concerning genetic testing and screening in public health. *American Journal of Medical Genetics, 125C*, 66–70.

Horstman, J. (2010). *The scientific American brave new brain*. San Francisco: Jossey-Bass.

Huizenga, H., van Zuijen, T. L., Heslenfeld, D. J., & Molenaar, P. C. (2001). Simultaneous MEG and EEG source analysis. *Physics in Medicine and Biology, 47*, 1737–1751.

Jaspers, K. (1949). *Vernunft und Existenz*. Bremen: Johs Storm Verlag.

Jotterand, F. (2008a). Beyond therapy and enhancement: The alteration of human nature. *Nanoethics: Ethics for Technologies that Converge at the Nanoscale, 2*, 15–23.

Jotterand, F. (2008b). *Emerging conceptual, ethical and policy issues in bionanotechnology*. Dordrecht: Springer.

Jotterand, F. (2010). Human dignity and transhumanism: Do anthro-technological devices have moral status? *The American Journal of Bioethics, 10*, 45–52.

Jotterand, F., & Giordano, J. (2011). Transcranial magnetic stimulation, deep brain stimulation and personal identity: Ethical questions, and neuroethical approaches for medical practice. *International Review of Psychiatry, 23*(5), 476–485.

Jotterand, F., Pascual, J. M., & Sadler, J. Z. (2013). The *can*'*t* and *don*'*t* of psychopathy: neuroimaging technologies, psychopaths and criminal responsibility. In J. Giordano (Ed.), *Neuroethics: Issues at the intersection of neuroscience and society*. Cambridge: Cambridge University Press (in press).

Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: Evidence for paralimbic system dysfunction. *Psychiatry Research, 142*, 107–128.

Kim-Cohen, J., Caspi, A., Taylor, A., et al. (2006). MAOA, maltreatment, and gene-environment interaction predicting children's mental health: New evidence and a meta-analysis. *Molecular Psychiatry, 11*, 903–913.

Klitzman, R. (2006). Clinicians, patients, and the brain. In J. Illes (Ed.), *Neuroethics: Defining the issues in theory, practice, and policy*. Oxford: Oxford University Press.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*, 673–676.

Larsson, H., Tuvblad, C., Rijsdijk, F. V., Andershed, H., Grann, M., & Lichtenstein, P. (2007). A common genetic factor explains the association between psychopathic personality and antisocial behavior. *Psychological Medicine, 37*, 15–26.

Matravers, M. (2010). In L. Malatesti & J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry, and philosophy*. Oxford: Oxford University Press.

McBride, D., & Schmorrow, D. (2005). *Quantifying human information processing*. Lanham, MD: Lexington Books.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Paris: Gallimard.

Mesulam, M. M. (2000). *Principles of behavioral and cognitive neurology* (2nd ed.). New York: Oxford University Press.

Meyer-Lindenberg, A., Buckholtz, J., Kolachana, B., et al. (2006). Neural mechanisms of genetic risk for impulsivity and violence in humans. *Proceedings of the National Academy of Science, 103*, 6269–6274.

Morse, S. J. (2010). Psychopathy and the law: The United States experience. In L. Malatesti & J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry, and philosophy*. Oxford: Oxford University Press.

Neumann, C. S., & Hare, R. D. (2008). Psychopathic traits in a large community sample: Links to violence, alcohol use, and intelligence. *Journal of Consulting and Clinical Psychology, 76*, 893–899.

Niedermeyer, E., & da Silva, L. (2004). *Electroencephalography: Basic principles, clinical applications, and related fields* (5th ed.). Philadelphia: Lippincott Williams & Wilkins.

Ogloff, J. P., & Wood, M. (2010). The treatment of psychopathy: Clinical nihilism or steps in the right direction? In L. Malatesti & J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry, and philosophy*. Oxford: Oxford University Press.

Ong, J. (2008). Deep brain stimulation: The quest for cognitive enhancement. *The Triple Helix, 5*, 6–8.

Perry, J. (2009). Diminished and fractured selves. In D. J. H. Mathews, H. Bok, & P. V. Rabins (Eds.), *Personal identity & fractured selves*. Baltimore: The Johns Hopkins University Press.

Plischke, H., Rousseau, D. D., & Giordano, J. (2012). EEG-based neurofeedback: The promise of neurotechnology and need for neuroethically-informed guidelines and policies. *J Ethics Biol Engineer Med, 4*, 7–18.

Pridmore, S., Chambers, A., & McArthur, M. (2005). Neuroimaging in psychopathy. *The Australian and New Zealand Journal of Psychiatry, 39*, 856–865.

Prinz, J. (2005). A neurofunctional theory of consciousness. In A. Brook & K. Akins (Eds.), *Cognition and the brain – The philosophy and neuroscience movement*. New York: Cambridge University Press.

Puranik, D. A., Joseph, S. K., Daundkar, B. B., & Garad, M. V. (2009). Brain signature profiling in India. Its status as an aid in investigation and as corroborative evidence – as seen from judgments. In *Proceedings of XX all India forensic science conference* (pp. 815–822), November 15–17, Jaipur.

Quednow, B. B. (2010). Ethics of neuroenhancement: A phantom debate. *BioSocieties, 5*, 153–156.

Raine, A., & Yang, Y. (2006). The neuroanatomical bases of psychopathy: A review of brain imaging sindings. In C. J. Patrick (Ed.), *Handbook of psychopathy*. New York: The Guilford Press.

Renaud, P., et al. (2011). Real-time functional magnetic imaging – brain-computer interface and virtual reality: promising tools for the treatment of pedophilia. In A. M. Green et al. (Eds.), *Progress in brain research* (pp. 263–272). Elsevier: Amsterdam.

Ridley, M. (2003). *Nature via nurture: genes, experience and what makes us human*. London: Harper Collins.

Rota, G., et al. (2008). Self-regulation of regional cortical activity using real-time fMRI: The right inferior frontal gyrus and linguistic processing. *Human Brain Mapping, 30*(5), 1605–1614.

Salekin, R. T. (2002). Psychopathy and therapeutic pessimism: Clinical lore or clinical reality? *Clinical Psychology Review, 22*, 79–112.

Shook, J., & Giordano, J. (2013). Toward a principled and cosmopolitan neuroethics. *Philosophy, Ethics and Humanities in Medicine* (in press).

Sitaram, R., Caria, A., Veit, R., Gaber, T., Rota, G., Kuebler, A., & Birnbaumer, N. (2007). fMRI brain-computer interface: A tool for neuroscientific research and treatment. *Computational Intelligence and Neuroscience*, 1–10, article ID 25487.

Sitaram, R., Caria, A., & Birnbaumer, N. (2009). Hemodynamic brain-computer interfaces for communication and rehabilitation. *Neural Networks, 22*, 1320–1328.

Vaadia, E., & Birbaumer, N. (2009). Grand challenges of brain computer interfaces in the years to come. *Frontiers in Neuroscience, 3*, 151–154.
van Outsem, R. (2011). The applicability of neurofeedback in forensic psychotherapy: A literature review. *The Journal of Forensic Psychiatry & Psychology, 22*, 223–242.
Viding, E., Blair, R. J., Moffitt, T. E., & Plomin, R. (2005). Evidence for substantial genetic risk for psychopathy in 7-year-olds. *Journal of Child Psychology and Psychiatry, 46*, 592–597.
Widiger, T. A. (2006). Psychopathy and DSM-IV psychopathology. In C. J. Patrick (Ed.), *Handbook of psychopathy*. New York: The Guilford Press.
Wurzman, R., & Giordano, J. (2012). Differential susceptibility to plasticity: A "missing link" between gene-culture co-evolution and neuropsychiatric spectrum disorders? *BMC Medicine, 10*, 37.
Yang, Y., Raine, A., Colletti, P., Toga, A. W., & Narr, K. L. (2009). Abnormal temporal and prefrontal cortical gray matter thinning in psychopaths. *Molecular Psychiatry, 14*, 561–562.

Section X

Ethical Implications of Sensory Prostheses

Ethical Implications of Sensory Prostheses

48

Sven Ove Hansson

Contents

Abstract

This survey begins with an overview of currently available and foreseeable sensory prostheses. Cochlear implants are now a routine technology for patients with a dysfunctional inner ear but a functional auditory nerve. Auditory brainstem implants are available for patients whose auditory nerve cannot be used. Visual prosthesis for the blind is a highly active research area but still far from results that can be used in routine clinical practice. Experiments are also made with artificial proprioception and touch for sensory feedback in limb

S.O. Hansson
Division of Philosophy, Royal Institute of Technology (KTH), Stockholm, Sweden
e-mail: soh@kth.se

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_46,

prostheses. Artificial biosensors are used in pacemakers, and research is being done on implantable drug delivery systems with biosensors that determine dosage. A wide range of ethical issues arise in connection with experiments and clinical usage of sensory prostheses: animal experimentation; informed consent, for instance, in patients with a locked-in syndrome that may be alleviated with a sensory prosthesis; unrealistic expectations of research subjects testing new devices; privacy issues for electronic implants with memory; security issues; effects of sensory improvements on a patient's personality and self-image; cultural effects of the new technologies in disabled communities; and the psychological and social effects of sensory enhancement.

Introduction

Neither the concept of perception nor that of a sense is precisely defined. We tend to use the word "sense" primarily about those of the registration mechanisms in our bodies that give rise to signals reaching our consciousness. This applies to the five classical human senses – sight, hearing, taste, smell, and touch – but also to several other mechanisms that provide us with information about the outside world or about the body itself. Our movements are guided by the sense of balance that is located in the vestibular system of the inner ear and also by proprioception that is based on sensory organs in muscles and joints. Our skin contains temperature receptors for cold and warm, and there are pain receptors both in the skin and on various internal surfaces of the body. We also have "internal senses" that register the body's homeostasis and the status of some of its organs. Chemoreceptors register the concentrations of oxygen, carbon dioxide, and a large number of hormones and other substances. Some of these registrations produce conscious sensations; for instance, receptors sensitive to insulin and nutrients in the blood have an influence on hunger. However, most of the chemoreceptors operate on an unconscious level. Here, the term "sense" will be used in a broad sense that covers sensors and receptors whose signals seldom if ever reach consciousness.

Artificial devices that improve, replace, or supplement our senses differ in the degrees to which they are integrated into the body. A magnifying glass is used to improve sight and a blind person's stick to partly replace it. We think of these as tools. Sense-enhancing tools such as eyeglasses and hearing aids that are worn for long periods of time are often referred to as assistive technologies. Assistive devices are called prostheses if they replace a body part and implants if they are operated into the body. Here, the focus will be on devices that are integrated with the body to a high degree, mainly prostheses and implants.

The purpose of this chapter is to provide an overview over sensory reconstructions and enhancements and the ethical issues that they give rise to. Paragraph 2 summarizes the status of current technology and technological development. Paragraph 3 introduces the major ethical issues that these technologies tend to give rise to. The chapter is followed by three others that provide in-depth discussions of the

most discussed ethical issues in this field, namely, cochlear implants (▶ Chaps. 49, "Ethical Issues in Auditory Prostheses"; ▶ 50, "Ethical Issues in Cochlear Implantation") and sensory enhancement (▶ Chap. 51, "Sensory Enhancement").

Currently Available and Foreseeable Devices

Hearing

Hearing aids in the form of ear trumpets have been used since the seventeenth century. Wearable electronic hearing aids have been available since the early twentieth century, and their development was a driving force in the early stages of electronic miniaturization (Mills 2011). However, hearing aids can only be used by persons whose hearing loss is not complete. A large technological step forward was taken with the introduction of cochlear implants. These are sound-activated devices that directly stimulate the auditory nerve, thereby bypassing dysfunctional parts of the inner ear. The first experiments with cochlear implants were made in the late 1950s, and they have been used clinically since the 1980s. Cochlear implants give rise to a sense of hearing that is sufficient for understanding speech. The best results are obtained when a deaf child receives the implants before 2 years of age (Peterson et al. 2010).

More recently, auditory brainstem implants have been developed for use in patients with a nonfunctional auditory nerve. The technology is largely the same, the major difference being the placement of the implant (Schwartz et al. 2008).

Seeing

Prosthetic vision for the blind ("bionic eyes") is a major research area in medical technology. The basic construction principle is the same as for cochlear implants, namely, that stimuli from technological sensors are relayed to the nervous system via a nerve-implant interface. The potential sites for implantation follow the nerve pathway that takes visual impressions to the brain: the retina, the optic nerve, the lateral geniculate nucleus, and the visual cortex (Rizzo et al. 2007). Retinal prostheses have the advantages of requiring less complicated surgery than the other options. They also make use of the natural processing system for preparing visual information before it reaches the cortex. However, the retina can only be used as an implant site in patients who have a large number of intact retinal ganglion cells (the cells that transmit visual impulses to the brain). Therefore, one of the other implant sites will have to be used, for instance, in patients with glaucoma or diabetic retinopathy (Fernandez and Hoffmann 2011).

Research on visual prostheses has been conducted since the 1960s (Rizzo et al. 2007). In 2011 there were about 23 devices under development, and five different research groups performed clinical trials with retinal implantation (Fernandez and Hoffmann 2011). A typical bionic eye consists of one implanted

and one external part. The implanted part is placed in the eyeball. The external part includes a small camera that is mounted on eyeglasses and a battery carried elsewhere on the body. Signals from the camera are sent wirelessly to the implant, where they are distributed to up to 60 electrodes placed on the retina, for instance, in a 6×10 array (Humayun et al. 2012).

Thus far, the results from retinal implants are much less impressive than those obtained with cochlear implants. Blind people obtain a black and white, coarse-grained "pixelated" impression of the environment. In the most successful cases, they can recognize a large letter or track a moving dot on a computer screen (Sample 2013). The usefulness of presently available bionic eyes to improve the quality of life has not yet been fully demonstrated (Fernandez and Hoffmann 2011).

Taste and Smell

Technologies are available that partially imitate human olfaction and taste ("electronic noses" and "electronic tongues"). However, currently this work has its focus on external measurement instruments (Wilson and Baietto 2009). Implantable devices seem to be rather far away from realization.

Proprioception and Touch in Limbs

Previously, limb prostheses were simple constructions like a wooden leg, and their function was primarily cosmetic and/or supportive. Today, myoelectric prostheses are being built that contain motors directly operated by the patient through connections with the nervous systems (Micera 2010). In order for these prostheses to function properly, sensory feedback is needed. Two types of sensors have been developed for such feedback, corresponding to touch respectively proprioception. Accelerometers have turned out to provide a useful mechanism for proprioception although they operate in a quite different way than natural proprioception. Accelerometer technology has also been developed for gait rehabilitation in hemiparetic patients. In the latter case, wearable accelerometers provide feedback to the patient, for instance, through sounds (LeMoyne et al. 2009).

In hand prostheses, touch sensors are used in combination with artificial proprioception. Both types of sensors can be connected either to a cortical implant or to nerves in the stump. It has been proposed that in the latter case sensory signals from the prosthesis can be connected to the remaining nerves in a way that corresponds to the patient's "phantom" sensations from the lost limb (Kroeker 2011).

The ultimate goal of connecting sensors in a limb prosthesis to the nervous system is to obtain a motor-sensory feedback loop similar to that of natural limbs. However, the difficulties in achieving this are tremendous, and despite considerable progress the functionality of currently available prostheses is still rather limited (Micera 2010).

Biomonitoring

A large number of research projects aim at the introduction of biosensors into the body in order to replace or supplement the body's own sensors. Such sensors may, for instance, continuously record electric signals in the brain or in the heart in order to detect dangerous events in these organs. They can also be used to follow pulse, blood pressure, intraocular pressure, or other physiological parameters. Sensors that measure bladder fullness can help patients with spinal cord injury to regain bladder control (Chew et al. 2013). Much of the research has been devoted to chemical sensors that register chemical health indicators such as the concentration of glucose and other metabolites, hormones, or drugs in the bloodstream or in specific organs (De Venuto and Vincentelli 2013; Vaddirajua et al. 2010).

Biomonitoring can be used either as a decision aid or as part of automatic feedback systems. As decision aids they provide the opportunity to follow diagnostic parameters continuously on all days. In addition they make it possible to collect diagnostic data at a distance. By holding, for instance, a mobile phone over the implant, the patient can transfer the data and send it to the clinic (Bauer 2007).

The use of biosensors in automatic feedback systems opens up new therapeutic prospects. Dynamic artificial pacemakers are already in widespread clinical use. They adjust the heart rate to physiological parameters, in a way intended to mimic the normal regulation of the heartbeat. The same technology can be used in mechanical circulatory support devices (left ventricle assist devices) and in future artificial hearts. Implantable cardioverter-defibrillators (ICDs) detect cardiac arrhythmia and correct it automatically by emitting an electric pulse (Puri et al. 2013).

Much research is conducted on using chemical sensors as parts of implanted drug delivery devices. This technology has the advantage of allowing for more precise and timely dosing. Furthermore, if the drug is delivered to the specific site where it is needed, higher local doses can be used with less systemic side effects than if the drug is distributed through ingestion or intravenous injection. In other words, this technology can shift the balance between therapeutic and adverse effects in a favorable direction.

Diabetes management is a particularly interesting application for implantable biosensors. Implanted glucose sensors can potentially improve the patient's monitoring of her disease, thereby improving insulin dosage. Future artificial pancreases will have to be connected to sensors that regulate insulin release (Srivastava et al. 2011). However, due to the body's reactions to implanted objects, it has turned out to be extremely difficult to construct glucose sensors that function reliably more than a week or so after the implantation. Inflammation and foreign body responses are still a major obstacle in the development of biosensors (Vaddirajua et al. 2010).

Enhancement and New Senses

Some of the technologies that have been developed for sensory prostheses can also be used for the sensory enhancement, i.e., they can make it possible to perceive

signals that a normal human being cannot perceive: infrared or ultraviolet light, ultrasound, infrasound, magnetic fields, or chemicals that are odorless for a normal human. Most such enhancements are far away from technical realization, but some low-tech enhancement technologies are already in use in the biohacking and body modification subcultures.

Probably the most common such implants are subdermal magnets, i.e., small silicon-coated magnets that are introduced permanently under the skin, usually on finger pads. Implanted magnets make it possible to feel certain electromagnetic fields, and some persons with these implants report that they can feel whether certain security alarms are on or off. The operation is offered commercially by some body modification practitioners, and reportedly thousands of people already have implanted magnets (Hameed et al. 2010; Firger 2013). Other types of sensors have been implanted in single individuals. For example, one person had a “third eye” in the form of a camera attached to implants at the back of his head. The camera took a photo every 60 s that was published on a website (Parry 2011). Another member of the body modification movement has a fairly large electronic device implanted under the skin of his forearm. Its function is to continuously record his body temperature and store this data (Woollaston 2013). An implantable compass is under development (Firger 2013).

Ethical Aspects

With the exception of cochlear implants, therapeutic sensory prostheses have not been subject to much ethical discussion. This may be because the purpose of such devices appears to be ethically unproblematic. But close inspection will show that quite a few ethical concerns need to be taken into account. Several of these concerns are common to the different types of sensory prostheses.

Animal Experiments

Like many other medical innovations, new implants are tested on animals before being tried out on humans. Animal experiments are ethically problematic, even when performed to avoid even more problematic testing on humans. It is generally agreed that the use of animals with higher taxonomical classification, in particular primates, should be avoided as far as possible. However, the function of implants in the nervous system is often difficult to test on the lower species, and therefore monkeys have often been used in these experiments. It is particularly important in such experiments to apply the so-called 3 Rs (Russell et al. 1959): replacement by nonanimal experiments as far as possible, for instance, through efficient use of simulation methods; reduction of the number of animals; and refinement by careful veterinary attention to reduce pain and distress (Sughrue et al. 2009; Gerrek 2009; Nobis 2009).

Informed Consent and Professional Responsibility

The implantation of sensory devices requires the informed consent of the patient. In some cases, informed consent may be difficult to obtain, or its quality can be questioned. This applies to people who are unable to communicate, for instance, due to a locked-in syndrome that may be alleviated with a sensory prosthesis. Another problem is that a patient's willingness to have an implant may be driven by unrealistic hopes about its efficiency. Since informed consent requires that the patient has understood the information, it may be questionable in such cases whether the patient's consent is fully informed. This seems to be a particularly serious problem in trials with bionic eyes. Due to the rudimentariness of the perception obtained with this technology, very few of the eligible patients are willing to volunteer as experimental subjects in implant trials. The few who volunteer tend to have unrealistic hopes about the outcome, hopes that cannot be fulfilled with the present technology (Xia and Ren 2013). It is essential that potential participants in trials with these and other sensory improvement technologies are sufficiently well informed and that they have a realistic understanding of the expected effects of the intervention.

According to the standard view in medical ethics, informed consent (when obtainable) is a necessary but not a sufficient condition for an intervention to be ethically acceptable. A reasonably favorable risk-benefit assessment is an independent ethical requirement and a sine qua non for an intervention to be acceptable. It is a violation of medical ethics to perform an intervention that does harm to the patient (without outweighing medical benefits), even if the patient consents to it and perhaps entreatingly asks for it. This is a pertinent issue not only in scientific experiments with sensory prostheses but also in relation to some of the body modifications undertaken by biohackers.

The abovementioned implantation of temperature-recording device into the body of a biohacker provides a clear example. No physician was willing to perform the surgery, presumably due to the potential risks involved in such an operation. Instead it was performed by a "flesh engineer" (Woollaston 2013). This raises an important ethical issue. If it would be unethical for a surgeon to perform an operation due to its high risks, can it be ethical for someone else such as a "flesh engineer" to perform it? This is part of a much larger issue, namely, whether the ethics of a profession such as the medical profession should be seen as the ethics of a particular class of activities or as the ethics of a particular class of persons. According to the former interpretation, the "flesh engineer's" operation would qualify as at least as unethical as the same operation hypothetically performed by a qualified surgeon. (Probably it should, under this assumption, be seen as even more unethical since lack of equipment and competence made it more risky than if a surgeon had performed it.)

Privacy and Security

The continuous monitoring of health-related parameters that is possible with implanted biosensors has an obvious and important advantage: It can be used to improve treatment. However, some of the information that is obtainable in this way

may be privacy sensitive, for instance, information about concentrations of alcohol, cotinine (a biomarker for tobacco exposure), or other substances related to recreational drug use or to compliance with lifestyle recommendations. Future multifunctional sensors can make such monitoring possible without the patient's knowledge or consent. Such monitoring could potentially lead to decisions contrary to the patient's interests, for instance, concerning eligibility for insurance.

There is also a risk of unauthorized access to information from implanted biosensors. Scannable implants can be read without the patient's consent, or the information contained from them can be intercepted when sent electronically from the patient to the healthcare provider (Bauer 2007; Bramstedt 2005). Like other digitized information, the information recorded in sensory devices can easily be stored. An electronic eye might, for instance, be provided with a memory function that saves everything it records. This may threaten the privacy of persons within camera reach of the person who carries the implant. When the implant carrier dies, the recorded data is still there and could potentially be archived against the person's wish.

In some cases it can be medically motivated to combine implanted biosensors with a geographical tracking unit (GPS receiver). If a biosensor detects a serious cardiac event or a severe hypoglycemia in a diabetic, an alarm signal to an emergency medical service, including the patient's geographical position, may save her life. But on the other hand, an implanted geographical tracking unit can be perceived as a serious infringement of privacy. It has been argued that the connection of implantable devices to external electronic communication networks may have large effects on how we see the relationship between the human body and its social and technological environment. Thus, "as we transform the human body internally with microchips and biosensors, we also transform externally how individuals interact and live in the world" (Bauer 2007).

The security of pacemakers, defibrillators, drug delivery devices, and other potentially lifesaving implants against unauthorized interventions is a serious matter that has not received sufficient attention. These devices can often be reprogrammed wirelessly, which is of course convenient in the clinic. However, some such devices are susceptible to reprogramming attacks that are performed with the intent to injure the patient (Halperin et al. 2008).

Effects on Personality and Self-Image

An individual's personality depends in part on her bodily experiences. New sensory experiences can potentially have effects on her personality and self-image. Although very little is known about such effects, the development of more efficient artificial biosensors gives us reason to watch out vigilantly for them. What effects can visual stimuli from a bionic eye have on a person with no previous experience of visual stimuli? What will the effects be of continuous exposure to new types of perceptual information, such as ultrasound or infrared light? Is there a risk of mental overload, or will there be personality changes?

There has been considerable debate about the phenomenological nature of the experiences obtained through sensory substitution, for instance, when "visual" information is mediated to a blind person through tactile stimuli. According to the so-called deference thesis, the experiences of such stimuli are visual in nature. According to the dominance thesis, these experiences remain within the tactile realm (De Preester 2011). Some authors have argued in favor of a third position, namely, that these sensations belong to a new category that can be subsumed under neither the replaced nor the replacing sensory modality (Farina 2013). This third position may be particularly plausible when the replacing sensory impressions are of an entirely new type, for instance, sensations obtained with electrical stimulation of the cortex. Although usually not discussed in ethical terms, the nature of "new" sensory modalities has ethical implications. The introduction of entirely new modes of perception would seem to increase the need for carefully following the psychological effects on the persons receiving these types of sensory impressions.

It is usually assumed that in order to be successful, a prosthesis has to be perceived by the patient as a part of her own body. However, in many cases the prosthesis has instead been perceived as a foreign object. Much can be learned from the experience of hand transplants. The recipient of the first hand transplant never accepted the new hand, and after 2 years he asked to have it amputated. Other patients have accepted a new hand but only after 4–5 months when sensation had been regained (Swindell 2007; Dickenson and Widdershoven 2001). Trials with artificial hands confirm the importance of the sensory component of advanced limb prostheses. The chances for a prosthesis to be effectively used and to be perceived as part of one's own body seem to increase with the introduction of realistic sensory feedback (Micera 2010).

Social and Cultural Effects

Medical interventions, including prostheses and implants, have effects not only on the individuals on which they are performed but also on social groups and on society as a whole. Radical improvements in treatment options will change the situation of disabled subcommunities in our societies. However, not all therapeutic improvements have been received positively in these subcommunities. The "fat is beautiful" movement denies that obesity is a disease requiring treatment and medical attention. Segments of the dwarf community have reacted against the introduction of therapies against their condition, seeing this as a threat to the future existence of their way of life and their organizations (Berreby 1996). By far the strongest such counterreaction is that of a part of the Deaf World against cochlear implant surgery in prelingually deaf children.

The criticism of cochlear implantation is strongly connected to a positive view of deafness in the Deaf World. Some of its members deny that they have an impairment or disability. Instead, they view themselves as members of a minority culture with its own language, customs, attitudes, knowledge, and values. Widespread use of cochlear implants will lead to a drastic decline in the population of this minority

culture. Deaf activist have often referred to the ethical principle that minority cultures should be preserved. They claim that large-scale implantation of children conflicts with the right of the deaf language and cultural minority to exist and flourish. The term "genocide" has sometimes been used to describe that prospect (Lane and Bahan 1998). This claim has given rise to an interesting discussion about the definition of a minority culture and whether cultures have intrinsic value (Levy 2002). Critics have pointed out the problematic nature of arguments that give precedence to the preservation of a culture over the interests of individual children. Some have noted that it is difficult to draw the line if cochlear implants are disallowed for this reason. If cochlear implants are unethical, then how should we judge the rubella vaccine (Balkany et al. 1996)? In clinical medicine, cochlear implants are commonly seen as an adequate treatment of deaf children, contrary to the Deaf World view just referred to. The reader is referred to the chapters authored by Thomas McCormick and the one from Linda and Paul Komesaroff for two nuanced views on this controversy, one with its roots in the medical community and the other in the Deaf World. (See also Clausen 2009.)

Enhancement may give rise to additional issues related to cultures and subcultures. It is not inconceivable that persons who have received certain enhancing interventions may come to form new subcultures. Furthermore, enhancement may change our views of normality, so that some of the unenhanced may come to be seen as "subnormal" in the relevant respect. If some people submit to enhancement, others may feel a pressure to follow suit, for the same reason that athletes who use performance-enhancing drugs induce others to do the same.

Some of the personal characteristics that can be enhanced through implantation or other interventions seem to have a function similar to that of positional (economic) goods, i.e., goods that give their owner a place in the social hierarchy. As an example of the latter category, having a color TV at the time when this was a new and exciting technology contributed positively to the owner's social status. This effect decreased in importance as color TVs became common. Access to a particular type of positional goods typically increases with economic growth, and it can then lose its positional value and be replaced by other objects as markers of social status. Personal characteristics with similar properties as positional goods may be called positional characteristics. Height is an example. If some of those predisposed to short stature are treated, the relative position of those untreated or untreatable can be expected to worsen. Experience from cosmetic surgery corroborates this mechanism. Our concepts of normal looks have changed when enhancing technologies became available. Hence, buckteeth were more accepted 50 years ago when orthodontic treatment was not available. The introduction of breast implants has had similar effects, and surgery that erases certain facial features has the same potential (Goering 2003).

Generally speaking, enhancement of a positional characteristic can have negative effects in the long run on the social situation of untreated individuals. Therefore, a proactive discussion is needed of distributive and procedural justice in relation to such enhancements. Cognitive and sensory capacities are examples of characteristics that may be positional.

On the other hand it must be conceded that progress is often hard to achieve unless some humans are allowed to go forward before others. At the bottom line, the enhancement issue is about what kinds of human beings there should be. Should future people be stronger and more intelligent than we are, and should they be able to directly perceive features of the environment that are inaccessible to us? Issues about what kind(s) of persons there should be are among the most difficult ones to deal with rationally in moral philosophy (Clausen 2008). The very basis for the discussion is insecure. What criteria should we use? Should we judge future persons by our own criteria or by the criteria that we predict (and partly determine) them to have? Possibly, the best way to tackle issues of enhancement is to do it incrementally, i.e., deal with each issue when it arises rather than trying to develop general principles that cover all potential future interventions in the human body.

Conclusion and Future Directions

We are probably only at a very early stage in the development of sensory prostheses. The experience so far gives us reason to believe that sensory prostheses can contribute substantially to people's quality of life, but we have also seen that the introduction of sensory prostheses can have negative psychological and social effects. Since the issues are largely similar for different types of sensory prostheses, we need to combine the rather fragmented ethical discussions on the different types into a more general, proactive discussion on the future of artificial sensory devices.

Cross-References

- ▶ Brain–Machine Interfaces for Communication in Complete Paralysis: Ethical Implications and Challenges
- ▶ Ethical Implications of Brain–Computer Interfacing
- ▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
- ▶ Extended Mind and Identity
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Neuroethics and Identity
- ▶ Neuroenhancement
- ▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Balkany, T., Hodges, A. V., & Goodman, K. W. (1996). Ethics of cochlear implantation in young children. *Otolaryngology – Head and Neck Surgery, 114*, 748–755.

Bauer, K. A. (2007). Wired patients: Implantable microchips and biosensors in patient care. *Cambridge Quarterly of Healthcare Ethics, 16*, 281–290.

Berreby, D. (1996). Up with people: Dwarves meet identity politics. *New Republic, 214*(18), 14–19.

Bramstedt, K. A. (2005). When microchip implants do more than drug delivery: Blending, blurring, and bundling of protected health information and patient monitoring. *Technology and Health Care, 13*, 193–198.

Chew, D. J., Zhu, L., Delivopoulos, E., Minev, I. R., Musick, K. M., Mosse, C. A., Craggs, M., Donaldson, N., Lacour, S. P., McMahon, S. B., & Fawcett, J. W. (2013). A microchannel neuroprosthesis for bladder control after spinal cord injury in rat. *Science Translational Medicine, 5*(210), 210ra155.

Clausen, J. (2008). Moving minds: Ethical aspects of neural motor prostheses. *Biotechnology Journal, 3*(12), 1493–1501.

Clausen, J. (2009). Man, machine and in between. *Nature, 457*(7233), 1080–1081.

De Preester, H. (2011). Technology and the body: The (im)possibilities of re-embodiment. *Foundations of Science, 16*, 119–137.

De Venuto, D., & Vincentelli, A. S. (2013). Dr. Frankenstein's dream made possible: Implanted electronic devices. In *Design, automation & test in Europe (DATE)* (pp. 1531–1536). Piscataway, NJ: IEEE.

Dickenson, D., & Widdershoven, G. (2001). Ethical issues in limb transplants. *Bioethics, 15*, 110–124.

Farina, M. (2013). Neither touch nor vision: Sensory substitution as artificial synaesthesia? *Biology and Philosophy, 28*, 639–655.

Fernandez, E., & Hoffmann, K.-P. (2011). Visual prostheses. In R. Kramme et al. (Eds.), *Springer handbook of medical technology* (pp. 821–834). Berlin: Springer.

Firger, J. (2013). The brave new world of biohacking. http://america.aljazeera.com/articles/2013/10/18/the-brave-new-worldofbiohacking.html

Gerrek, M. L. (2009). Primate stroke research: Still not interested. *American Journal of Bioethics, 9*(5), 29–30.

Goering, S. (2003). Conformity through cosmetic surgery: The medical erasure of race and disability. In R. Figueroa, S. Harding, & S. G. Harding (Eds.), *Science and other cultures: Diversity in the philosophy of science and technology* (pp. 172–188). New York: Routledge.

Halperin, D., Heydt-Benjamin, T. S., Ransford, B., Clark, S. S., Defend, B., Morgan, W., Fu, K., Kohno, T., & Maisel, W. H. (2008). Pacemakers and implantable cardiac defibrillators: Software radio attacks and zero-power defenses. In *Proceedings of the 2008 IEEE symposium on security and privacy* (pp. 129–142). Piscataway, NJ: IEEE.

Hameed, J., Harrison, I., Gasson, M. N., & Warwick, K. (2010). A novel human-machine interface using subdermal magnetic implants. In *IEEE 9th international conference on cybernetic intelligent systems (CIS)* (pp. 1–5). Piscataway, NJ: IEEE.

Humayun, M. S., Dorn, J. D., da Cruz, L., Dagnelie, G., Sahel, J.-A., Stanga, P. E., Cideciyan, A. V., Duncan, J. L., Eliott, D., Filley, E., Ho, A. C., Santos, A., Safran, A. B., Arditi, A., Del Priore, L. V., & Greenberg, R. J. (2012). Interim results from the international trial of second sight's visual prosthesis. *Ophthalmology, 119*, 779–788.

Kroeker, K. L. (2011). Engineering sensation in artificial limbs. *Communications of the ACM, 54*(4), 16–18.

Lane, H., & Bahan, B. (1998). Ethics of cochlear implantation in young children: A review and reply from a Deaf-World perspective. *Otolaryngology – Head and Neck Surgery, 119*, 297–313.

LeMoyne, R., Coroian, C., Mastroianni, T., Opalinski, P., Cozza, M., & Grundfest, W. (2009). The merits of artificial proprioception, with applications in biofeedback gait rehabilitation. Concepts and movement disorder characterization. In C. A. B. de Mello (Ed.), *Biomedical engineering* (pp. 165–198). Rijeka: InTech.

Levy, N. (2002). Reconsidering cochlear implants: The lessons of Martha's Vineyard. *Bioethics, 16*, 134–153.

Micera, S. (2010). Control of hand prostheses using peripheral information. *IEEE Reviews in Biomedical Engineering, 3*, 48–68.

Mills, M. (2011). Hearing aids and the history of electronics miniaturization. *IEEE Annals of the History of Computing, 33*(2), 24–45.

Nobis, N. (2009). Interests and harms in primate research. *American Journal of Bioethics, 9*(5), 27–29.
Parry, M. (2011, February 7). Health problems force professor to pull camera from back of head. *Chronicle of Higher Education*.
Peterson, N. R., Pisoni, D. B., & Miyamoto, R. T. (2010). Cochlear implants and spoken language processing abilities: Review and assessment of the literature. *Restorative Neurology and Neuroscience, 28*(2), 237–250.
Puri, M., Chapalamadugu, K. C., Miranda, A. C., Gelot, S., Moreno, W., Adithya, P. C., Law, C., & Tipparaju, S. M. (2013). Integrated approach for smart implantable cardioverter defibrillator (ICD) device with real time ECG monitoring: Use of flexible sensors for localized arrhythmia sensing and stimulation. *Frontiers in Physiology*, *4*(article 300).
Rizzo, J. F., III, Snebold, L., & Kenney, M. (2007). Development of a visual prosthesis. A review of the field and an overview of the Boston retinal implant project. In J. Tombran-Tink, C. Barnstable, & J. F. Rizzo (Eds.), *Ophthalmology research: Visual prosthesis and ophthalmic devices: New hope in sight* (pp. 71–93). Totowa: Humana Press.
Russell, W. M. S., Burch, R. L., & Hume, C. W. (1959). The principles of humane experimental technique. Available on http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc
Sample, I. (2013, February 20). Retinal implant restores partial sight to blind people. *Guardian*.
Schwartz, M. S., Otto, S. R., Shannon, R. V., Hitselberger, W. E., & Brackmann, D. E. (2008). Auditory brainstem implants. *Neurotherapeutics, 5*(1), 128–136.
Srivastava, R., Jayant, R. D., Chaudhary, A., & McShane, M. J. (2011). 'Smart tattoo' glucose biosensors and effect of coencapsulated anti-inflammatory agents. *Journal of Diabetes Science and Technology, 5*, 76–85.
Sughrue, M. E., Mocco, J., Mack, W. J., Ducruet, A. F., Komotar, R. J., Fischbach, R. L., Martin, T. E., & Connolly, E. S., Jr. (2009). Bioethical considerations in translational research: Primate stroke. *American Journal of Bioethics, 9*(5), 3–12.
Swindell, J. S. (2007). Facial allograft transplantation, personal identity and subjectivity. *Journal of Medical Ethics, 33*, 449–453.
Vaddirajua, S., Tomazos, I., Burgess, D. J., Jain, F. C., & Papadimitrakopoulos, F. (2010). Emerging synergy between nanotechnology and implantable biosensors: A review. *Biosensors and Bioelectronics, 25*, 1553–1565.
Wilson, A. D., & Baietto, M. (2009). Applications and advances in electronic-nose technologies. *Sensors, 9*, 5099–5148.
Woollaston, V. (2013, November 4). Now THAT'S 'wearable technology'! Man implants a mini computer under his SKIN to track his body temperature. *Daily Mail*.
Xia, Y., & Ren, Q. (2013). Ethical considerations for voluntary recruitment of visual prosthesis trials. *Science and Engineering Ethics, 19*, 1099–1106.

Ethical Issues in Auditory Prostheses

49

Thomas R. McCormick

Contents

Abstract

The innovation of cochlear implants and other auditory prostheses designed to improve hearing and oral communication for deaf and profoundly hard-of-hearing individuals is relatively new. Although such devices open the possibility of hearing, their use as a clinical treatment first for adults and later for children has been accompanied by a number of ethical conflicts and controversies. Compassionate physicians began to use this invention in both adults and children before scientific testing could be carried out to measure the risks and benefits of the new treatment. Controversy swirled around questions of the most appropriate age to implant cochlear devices in children. Early devices were extremely limited, and there were notable instances of device failure. The cost of this technology raised issues of cost containment in an era of scarce resources. The relatively rapid rise in the use of a treatment that allows deaf persons to hear has had a dramatic impact upon the Deaf community whose members fear a radical reduction in Deaf members, leading to

T.R. McCormick
Department Bioethics and Humanities, School of Medicine, University of Washington, Seattle, WA, USA
e-mail: mccormic@u.washington.edu; mccormic@uw.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_103,

a diminished community, fewer people using American Sign Language (ASL), and the loss of resources that have traditionally been available to promote processes of education and enhancement for those unable to hear. The judicious application of ethical principles offers promise in identifying ethically acceptable options toward resolving these issues. Ethical principles of beneficence and nonmaleficence require that we carefully examine issues of risk and benefit for patients. The principle of autonomy underscores the importance of informed consent and informs us of the duties of surrogates when deciding for minor children. Respecting diverse values within a pluralistic society places a strong emphasis on exercising principles of justice and fairness.

Introduction

A prosthesis is considered to be an artificial device used to replace a missing body part such as a limb, a heart valve or, to assist the functioning of an impaired organ. It is widely held that one's natural organs and limbs are superior to any synthetic replacement. Therefore, readers must have been startled at the title of a *Scientific American* article, asking "Should Oscar Pistorius's Prosthetic Legs Disqualify Him from the Olympics?" (Eveleth 2012). It was held by some that Oscar's prosthetic legs gave him an unfair advantage over other runners with two normal legs. Oscar suffered bilateral leg amputations as a child, and as an adult, began running with prosthetic legs called "Cheetahs." He was so successful in the Special Olympics that he decided to try out for the Olympics. In this context, a controversy broke out over possible advantages Oscar might enjoy by running with the prosthetic legs and feet. Eventually, Oscar was allowed to compete in the 2012 Olympics and ran hard, but did not win a medal. Still, many gained a new awareness that prosthetic legs and feet have undergone almost unimagined improvements in recent years. Similarly, improvements continue in the field of auditory prostheses. Today's cochlear implants are a vast improvement over the initial device implanted in the mid-1980s, and new types of implements, such as the hybrid, are currently in the testing stages. This chapter explores ethical issues related to the advent and use of auditory prostheses.

When it comes to auditory prostheses, even with major improvements, no one has yet made the argument that these synthetic devices are superior, or even nearly as good as the natural organs for hearing. However, there is a sense of amazement that after centuries of human existence in which little could be done to allow or enhance hearing for the deaf and the profoundly hard of hearing, great strides have been made in just the past few decades through the invention of hearing-aids and cochlear implants. These innovative devices, along with other inventions such as the auditory brainstem implant and vestibular-cochlear prosthesis, offer new hope for deaf individuals who are desirous of hearing and speaking as opposed to communicating primarily through American Sign Language (ASL).

Brief History of Deafness and Society's Responses

It is theorized that a small minority of persons were deaf from the earliest history of humankind. Deaf individuals were mentioned in ancient Jewish literature, and their special needs were recognized within that caring community. In Gannon's *Deaf Heritage*, several early historical references are made concerning deaf persons. In 355 BCE, Aristotle claimed that those "born deaf become senseless and incapable of reason" (Gannon 1981). Aristotle also claimed, "Men that are deaf are in all cases dumb; that is they can make vocal sounds, but they cannot speak" (Gannon 1981). Of course, we know today that Aristotle thought wrongly, having based his statements on his limited personal observations. Nearly 1,000 years later, in 721 AD, St. Bede wrote about St. John of Beverly teaching a deaf-mute to speak, a notable achievement (Gannon 1981). Rudolphus Agricola (1443–1485) writes about a deaf-mute who learned to read and write. Girolamo Cardano (1501–1576) is the first physician of record to recognize the ability of the deaf to reason (Gannon 1981). In 1661, two deaf persons, Matthew Pratt and Sarah Hunt, married in Weymouth, Massachusetts. Cotton Mather claimed that Sarah "spoke with Signs" and that her children learned to speak "sooner with eyes and hands than by their lips" (Lang 2007). Benjamin Franklin and John Quincy Adams were notable Americans who became aware of pioneering efforts to educate the deaf in the eighteenth century. In 1784, when Adams was ambassador to France, he sent a letter to William Cranch of Cambridge, Massachusetts, describing his impressions of the school in Paris of Abbe Charles-Michel de l'Epee. Adams wrote that Epee taught deaf pupils "not only to converse with each other by signs, but to read and write, and comprehend the most abstracted metaphysical ideas" (Lang 2007). In short, from earliest times, there have been references to the deaf and a gradual, growing appreciation in the hearing community for their capabilities. Anecdotally, deaf persons used hand-signs from earliest times, using simple, intuitive gestures to communicate their needs to others. In the 1500s, in Europe, sign language began to be organized as a formal communication system (Naff 2010). Deaf persons married, raised families, owned property, and functioned effectively in society.

In contrast to Aristotle's notions, as it became more widely accepted that deaf persons were indeed intelligent and capable of reasoning, more and more efforts to provide education of the deaf were undertaken and new means were utilized to assist those without hearing to communicate. Schools for the deaf sprang up in several countries. France had a school for the deaf. After studying techniques in educating the deaf in France, in "1816 Laurent Clerc, a deaf teacher of the deaf, and Thomas H. Gallaudet, a minister, traveled to the United States where they established the Connecticut Asylum for the Education and Instruction of Deaf and Dumb Persons, the first permanent school for the deaf in America, opened in Hartford "on April 15, 1817" (Crouch 2002). Their efforts contributed greatly to the spread of what would soon be known as American Sign Language (ASL). In 1864, Gallaudet's son, aided by Clerc and colleagues, established the Columbia Institute

for the Deaf in Washington D.C., which has since been renamed Gallaudet University and has served as an important institution of higher education for deaf persons (Naff 2010).

In the early years of deaf education, controversy arose over the most effective ways to teach methods of communication to deaf people. The oralists and the manualists represented two different ideas about communication. The oralists believed the deaf should be taught to speak and to lip-read. The manualists believed that deaf persons should be taught sign language. Alexander Graham Bell, inventor of the telephone, whose wife was deaf, believed strongly that learning to speak and to lip-read were important so that deaf persons could live and work within the majority hearing society. Others had far greater confidence in signing as a way for deaf persons to communicate. Gradually, a Deaf culture emerged and deaf persons who identified with the Deaf culture argued that deaf persons will never hear or speak like the hearing and so should abandon efforts to do so and become proficient in sign language. The debate continues to this day and as will be shown later on in this chapter, and as well in (▶ Chap. 50, "Ethical Issues in Cochlear Implantation") and gives rise to some of the ethical issues in the use of auditory prostheses.

In the last century, a number of assistive devices have come onto the market and many hearing-impaired individuals use these external assistive devices in their daily lives. Deaf and hard-of-hearing individuals can communicate by telephone using telecommunications device for the deaf (TDD). A hearing-impaired person can also communicate over the phone with a hearing person via a human translator. The use of internet, email, and mobile phone text-messaging are beginning to take over the role of the TDD. Software programs are now available that automatically generate a closed-captioning of conversations. Examples include discussions in conference rooms, classroom lectures, and religious services. Skype and similar video technologies can be used for distance communication using sign language. Video-conferencing technologies permit signed conversations as well as permitting a sign language English interpreter to voice and sign conversations between a hearing-impaired person and that person's hearing party. Other assistive devices include the use of flashing lights to signal events such as a ringing telephone, a doorbell, or a fire alarm. Hearing-dogs are a specific type of assistance dogs specifically selected and trained to assist the deaf and hearing-impaired by alerting their handler to important sounds, such as doorbells, smoke alarms, ringing telephones, or alarm clocks.

Hearing-aids have become more powerful and offer enhanced hearing to some who maintain a residual ability to hear sounds. Over 1,000 different models are available in the United States although they all include a microphone to pick up sound, an amplifier to boost sound, a speaker to deliver sound to the ear, and a battery power source. Over 65 % of users employ binaural hearing-aids. It is estimated that there are 28 million Americans who have "ski-slope" loss in which their ability to hear high-pitched sounds plummets radically (Nuzzo 2010). These patients are likely candidates for a newer "hybrid" cochlear implant that is currently being tested.

The Cochlear Implant

Without a doubt, the most notable innovation has been the invention and development of the cochlear implant device. The cochlear device has several parts, using an external microphone to pick up sounds from the environment and send them to a receiver implanted under the skin behind the ear which relays these to an array that has been carefully placed inside the cochlea. Signals from an array of about 22 electrodes are sent to the auditory nerve and then to the brain. With practice and assistance from an audiologist, the deaf person with a cochlear implant can learn to recognize the sounds of speech and can more successfully pick up spoken language, usually with the assistance of a speech therapist in the early stages as the brain is adapting to perception of sounds. Cochlear implant recipients often augment hearing by lip-reading. Cochlear implants are described by Hansson as a new and important class of implants interfacing with neural tissue to treat deafness (Hansson 2005). Graeme Clark, inventor-developer of the multiple channel electrode array, claims this prosthetic innovation is the first major device since sign language was developed over 200 years ago to assist profoundly deaf people in hearing speech and other sounds, enabling them to communicate more freely (Clark 2003).

On February 25, 1957, Dr. Andre Djourno and Dr. Charles Eyries, in France, were the first surgeons to implant a single electrode in the cochlea of a patient. Due to a prior surgery to remove large bilateral cholesteatomas, this 57-year-old male was left deaf and with facial paralysis. The primary purpose of the surgery was to provide a right-side facial nerve graft using fetal sciatic nerve tissue. Djourno had been experimenting with nerve-stimulating devices and suggested that since the site was exposed, there was little to be lost by implanting an electrode in the cochlea, and Eyries, the surgeon, agreed and both surgeries were carried out. Apparently, the nerve graft worked, and as well, the electrode allowed the patient to hear rudimentary sounds (Eisen 2006).

In the late 1960s, in the United States, Dr. William House worked with Jack Urban, a skilled engineer who had developed some excellent new neurotologic medical instruments. By this time, there was evidence that synthetic materials, such as pacemakers, could safely be implanted in the human body. Once reassured about this safety issue, House and Urban turned to the issue of efficacy. One patient in particular, Charles Graser, who had been deaf for 10 years from ototoxicity, was a willing subject for their experiments. They further developed the single-electrode implant and teamed with 3 M to manufacture these (called the House/3 M cochlear implant) in the United States in the 1970s (Eisen 2006).

Other pioneers in developing the cochlear implant were Robin Michelson, an otolaryngologist, and Michael Merzenich, a neurophysiologist, brought together as a team by Francis Sooy, chairman of a small Department of Otolaryngology at University of California, San Francisco (UCSF). Their work determined that in order to convey complex sounds such as speech, a multiple electrode array would be necessary (Merzenich 1973). The race to develop a multiple electrode array was inspired by the presentation of their work at the American Otological Society meeting in 1973.

In 1974, the NIH took the lead in requesting an objective assessment of the progress in cochlear implants. The contract was awarded to Robert Bilger and his team from the University of Pittsburgh. Thirteen subjects with cochlear implants were flown to Pittsburgh. Testing demonstrated that the single-channel electrode did not allow subjects to understand speech. However, it did help lip-reading, improved quality of life, and significantly helped the speech production of the subjects. Although the implantation of single electrodes was soon to come to an end, the Bilger Report, showing objective benefits and minimal risk, actually opened the door for funding research into a multiple channel device (Eisen 2006). Several thousand patients had been implanted with single-electrode devices by 1984, when the Food and Drug (FDA) granted formal approval of the device. It is clear that the clinical application of these implants prior to thorough studies on the risks and benefits of the new device and prior to FDA approval contributed to the ethical controversy of the time.

In about the same time period, Graeme Clark was experimenting with the development of a multielectrode cochlear implant. He hypothesized about the limitations of the single electrode as early as 1969 in his graduate thesis at the University of Melbourne (Clark 1969). Clark worked on developing all of the components for a multi electrode array. He implanted his device in an adult patient in 1978. As described by Lantos, "Multichannel devices divide the incoming signal into various frequency bands that are then transmitted to various sites of stimulation spanning the inner ear. Low-pitch sounds are sent to one part of the cochlea, high-pitch sounds to another, more closely mimicking the human ear. Because multiple channels provide a more detailed representation of sound, they are thought to allow better speech understanding than do single-channel devices. Clarke's device was approved in 1985" (Lantos 2012). Note, that although the FDA had not yet given approval for the use of CI in children, they were nevertheless used in the clinical context as therapy for deaf children. Many, particularly in the Deaf community, believed such applications to be unethical in the absence of clinical trials demonstrating their safety and efficacy in children. Finally, in 1990, the FDA approved the use of cochlear implants in children aged 2 and over (Institutes of Health 1990).

Ethical Issues

The FDA's decision to allow implants in children as young as 2 years had the effect of throwing fuel on the fires of controversy. Emotions flared and members of the Deaf community claimed that allowing cochlear implants in young children was an extreme threat to Deaf Culture (Lantos 2012). The following year, 1991, spokespersons for the National Association of the Deaf (NAD) issued a statement deploring the FDA's decision. They argued that cochlear implants in children were experimental, and studies had not been done to produce good evidence of usefulness. They noted that the FDA did not consult with experts on deafness and deaf education and that the value and benefits of sign language had not been fully considered. They claimed that the psycho-social development of deaf children

had not been fully considered. They demanded that surgeons discontinue implanting children until studies were completed. Most controversial was the claim of NAD that deafness “comprises a linguistic and cultural minority” and their protest that children should not be subjects of biological engineering that would change them from being a full member of a minority community and result in membership in the majority community. In addition, NAD recommended a national conference to address ethical issues surrounding cochlear implants in children (NAD 1991).

In response, a conference was convened at National Institutes of Health (NIH) in 1995, but it was not the conference that NAD had intended. The NAD wanted the conference to address the broad spectrum of ethical issues in cochlear implant. Instead of addressing these ethical issues, the conference had a more narrow focus on new innovations in cochlear devices and implant surgery and new information about the efficacy of such interventions in deaf persons. Although the conference was multidisciplinary in nature with representatives from otolaryngology, audiology, speech-language pathology, pediatrics, psychology, and education, and included a public representative, “there was no representative from Deaf culture” (Lantos 2012). Whether an oversight, or intended, an opportunity was lost for rapprochement.

The 1990s was tumultuous with both sides in the controversy seeming to dig-in to established positions. The Deaf culture claimed that not enough study had gone into examining and analyzing the efficacy of cochlear implants and that innovation and technology had unduly charmed the public into believing something as yet unproven. Some in the Deaf culture interpreted the use of cochlear implants as a direct assault on their Deaf community with the intent of minimizing the ranks of people communicating in sign language and “mainstreaming” more and more deaf individuals – thus reducing the population of the Deaf culture.

On the other side, many in the health care community felt that an innovative treatment that could help deaf persons hear, learn to speak fluent English, and improve their situation in life in terms of education, employment, and social relationships was being prohibited by the NAD for the sake of its own members. They argued that 95 % of deaf children are born to hearing parents who want their deaf children to hear with the assistance of the cochlear implant and to speak and participate in their family and community (McCormick 2010).

Just when it seemed this controversy was at a stand-off, there was a shift, primarily coming from the Deaf community and NAD. A new position paper was drawn up by NAD in 2000 that softened its earlier opposition and for the first time recognized cochlear implants as one of several options that were available to deaf persons. It encouraged parents to study the many diverse options open to themselves and their deaf infants, including information about ASL and the deaf community, so that they could render “informed consent” in a meaningful way when it became time to choose a treatment option (NAD Cochlear Implant Committee 2000).

Ethical issues continue. Current ethical issues include concerns over the safety and efficacy of cochlear implants, and whether sufficient research has been done to

provide scientific evidence of proportional benefit to those implanted with the cochlear device. Additionally, there is ethical controversy over the issue of device failure and the identification of faulty devices and whether they should be removed and replaced and if so, who should pay for such procedures? Controversy continues over the optimal age in children for cochlear implant. Since parents are surrogate decision-makers for minor children, how can decisions be formed that are in the "best interests" of the affected children when parents have beliefs that may delay or prohibit cochlear implant? In such cases, should health care professionals abandon the ethics of consensus-building and resort to force, for example, by notifying Child Protective Services (CPS) of possible medical neglect and asking the courts to intervene on behalf of minor children by mandating a cochlear implant? What about deaf parents who have deaf children and want their deaf child to learn ASL and to become a participating member of the Deaf community, how do we balance the duty of "respect" for a family's unique value system and our duty of "beneficence" to a dependent, minor child? On the other hand, some deaf patients have found their cochlear implant is sufficiently helpful that they now desire a second implant to provide bilateral hearing. In an era of scarce resources, should there be policies that encourage or restrict payment for bilateral cochlear implants?

Informed consent is a fundamental ethical principle in modern health care, stemming from the more basic principles of respect for persons and respect for patient's autonomy. Obtaining informed consent from a patient is a dynamic process. First, the patient must be competent and possess decisional capacity. Full information pertaining to any medical intervention must be communicated by the health care professional in a manner that is understandable to the patient. The proposed benefits and goals of treatment must be discussed, as well as the risks and factors that might be obstacles to the success of the proposed procedure. In this communicative process, the health care professional must be assured that the patient has both an understanding and an appreciation of all of the facts that are relevant to his or her situation. Further, the patient must voluntarily accept the proposed treatment, that is, without coercion (Beauchamp and Childress 1994). The process is usually concluded when the patient is properly informed and either rejects the treatment or chooses to proceed with the procedure and signifies this by voluntarily signing an "informed consent" document or operating permit. Unfortunately, there are still some who mistakenly believe that the important matter is obtaining the signed document. It cannot be stated strongly enough that informed consent is a dynamic process that the professional must conduct with respect for the unique situation of any particular patient so that such understanding and appreciation for risks and benefits is fully apprehended. Further, since physicians occupy a position of power within the society, care must be taken that the choice is freely made by the patient.

It has been suggested that since deafness is not a life-threatening condition, a family could justifiably delay any decision on cochlear implant until their deaf child became 18 years of age and could make the decision as an adult. Although this proposal is respectful of the autonomy of both the parents, and prospectively seems to respect the autonomy of the patient, it is problematic because too much time will

be lost between diagnosis and age 18. After age 7, the brain loses much of its plasticity and precious years of hearing and learning speech have been lost for the patient, perhaps irrevocably. In the light of the ongoing support from the parents that is needed by any child undergoing cochlear implant surgery and the family's commitment to support the child through speech and hearing therapies, it is not considered propitious to seek a court mandate authorizing such surgery against the wishes of the non-consenting parents.

Consent is a particularly complex issue when a cochlear implant is considered for a child. FDA in 2000 granted approval for cochlear implants in children as young as 12 months (NIH 2010). An adult can participate in the consenting process and freely elect to have a cochlear implant, or to decline, without incurring an ethical issue. However, decision-making authority in medical care for children resides with the parents. Parents, acting as surrogate decision-makers, must consent for the surgery that allows a device to be placed inside their child's cochlea. Parents have a moral duty to act in ways that further the best interests of their children, but it is quite understandable that parents have their own values and loyalties and may find it difficult to sort out a clear picture of which process will advance the best interests of minor children who are dependent upon them to make such decisions (Beauchamp and Childress 1994). The advantages of early implantation (between 12 and 18 months) are becoming increasingly clear. "The most impressive gains were demonstrated by the children implanted between 12 and 18 months of age. Waltzman reported that over half of the children in the youngest age group achieved auditory milestones approximating those of hearing peers after only 6 months of device use. For the children implanted between 19 and 23 months, only about one-fourth attained scores within the broad normal range after 6 months of device use" (Waltzman 2006).

The young brain is most plastic, and the brain's ability to hear the sounds of speech through the implant and to recognize sounds for speech formation is optimal at an early age. Children implanted at an early age have more incidental learning of tonal changes and inflections from hearing their parents speak that augment their training. Parents have a moral duty to act in the best interests of their affected children, or at least to avoid preventable harm. Is it not a harm to deprive a child from hearing and the optimal chance for verbalized communication? Conversely, some parents argue that it is a harm to implant a device in an operation that will cause some harm to cochlear tissue, increase the risk of meningitis, and remove the child from a life within the Deaf Community, thus forcing that child to participate and compete in a hearing world.

In the ethical debate, opponents of cochlear implants have cited a variety of risks for young patients such as, uncertainty of the diagnosis, increased risk for meningitis and otitis media, as well as the inevitable damage that is done to tissues in the cochlea in the process of inserting the electrodes. The FDA reported in 2002, their awareness of 91 reports worldwide of meningitis in patients implanted with 3 approved devices, resulting in 17 deaths (FDA 2002). The Center for Disease Control and Prevention (CDC) and the FDA along with State Departments of Health began an investigation to determine risk factors that lead to meningitis in

patients under the age of 6 at the time of implant (FDA 2002). In light of the discovery that there were more cases of meningitis with the CLARION device and positioner, Advanced Bionics agreed to halt the use of its positioner and voluntarily recalled any unimplanted CLARION devices in the United States. Neither the Cochlear Limited nor MedEl Corporation electrodes uses a positioner (FDA 2002). There remains a significant population of adults and children with positioners still in place as the FDA did not recommend prophylactic removal. This is a continuing controversial issue (McCormick 2010).

The risk of contracting meningitis has always existed for implant patients, but with the advent of better immunization, the risk levels continue to fall. In the light of increased risks to patients for contracting meningitis, it is concerning that immunization rates are considerably below the optimal level in transplant patients. This raises an ethical duty for all surgeons involved in cochlear implants to be sure that all implant recipients should be immunized. Further, cases of otitis media should be treated aggressively in an effort to minimize risks from infection that might progress to meningitis (Melton and Backous 2011).

Surgeons have an ethical duty to exercise great caution in the insertion of electrodes to minimize cochlear trauma. Pediatricians caring for deaf children contemplating cochlear implant have a duty to inform, recommend, and administer PCV-13 immunization to all cochlear candidates. Since the advent of the new 13-valent immunization vaccine in 2010, replacing the 7-valent version, it is predicted that there will be a continuing drop in infections among children receiving implants. CDC guidelines recommend the 13-valent immunizations for all children from 2 months to 18 years and, 8 weeks following completion of the PCV-13 series, a single dose of PPV-23 (Melton and Backous 2011). Melton and Backous claim that cochlear implants are currently the standard of care to optimize hearing and speech for children with deafness and that the driving forces supporting this are the scientific facts about the significance of early auditory development (Melton and Backous 2011).

The FDA currently approves cochlear implants to children with deafness down to year 1 (Institutes of Health). Some suggest that the age of implantation could be lowered safely to age 6 months. However, it appears that implantation between 12 and 18 months allows more time for an accurate diagnosis, testing for additional disorders, and greater development of skull thickness prior to implantation.

Considerations favoring early usage of cochlear implants stem from results of several studies showing that although the language and hearing results from cochlear implants were variable, in most cases, it provides a clear advantage. A recent study has reported that child and adolescent evaluation of overall quality of life are about the same between normal hearing children and those with cochlear implants (Loy 2010). The overall picture has formed that the benefits from cochlear implants outweigh the burdens. Gradually, in the United States, a broad consensus has emerged around the ethical duty to maximize the learning and communicating opportunities for deaf children through the use of all means available, or total communication (TC). In the United States, children use both ASL and oral communication. Hopefully, within Deaf culture, there will also be room for both.

Cochlear implants are now being employed in most developed countries in response to childhood deafness. A growing number of studies of this phenomenon have reported on the recipient's audition, speech perception and production, and the development of spoken language. One such study carried out in an eastern Australian study in 2008 involved parents and teachers of 1,260 children who had received cochlear implants prior to age 18. Researchers concluded that the results of the study are largely positive but cautioned that parents may be slower to accept that signed communications will be part of the lives of many of these children. While making clear the benefits of cochlear implants, the study suggested that ongoing efforts must be made to meet the challenges for children with implants and their families to assist in their social development and academic achievement. The study made clear that educators who work with these children need professional development and training about the CI devices and the special needs of children with implants to help meet this challenge (Punch and Hyde 2011). Dr. Lee, a deaf scholar, claims that "*all* implanted children should have the opportunity to learn sign language at a very early age, even if only as an alternate language" (Lee 2012).

Future Issues

New prosthetic devices will inevitably enter the market. Companies that make cochlear devices are continuously at work to develop more efficient implements. For example, a totally implantable cochlear implant could be made in the future. As research develops on auditory brainstem implants, patients who might benefit will be better served as safety and efficacy issues are resolved. Currently, there is a relatively small population who qualify for ABI. "An FDA-approved auditory brainstem implant (ABI) is considered medically necessary in an individual when all of the following criteria are met: 1. Is 12 years of age or older; and 2. Diagnosed with neurofibromatosis type 2; and 3. Is completely deaf as a result of bilateral neurofibromas of the auditory nerve" (Anthem Blue Cross-Blue Shield Medical Policy 2012).

Patients with both impaired hearing and balance function, a situation especially prevalent in China and India due to ototoxic drugs and genetic mutations, may be helped in the future by an integrated vestibular-cochlear implant. The theory and technology for electrical stimulation of vestibular nerves is established, but development has been slowed by the need for miniaturization and mass production of gyroscopic sensors. It is expected that these devices will be available in the future (Lu 2011).

It is inevitable that genetics will also play a role in ethical decisions pertaining to deaf persons. Since the publication of the human genome, the cost for decoding one's individual genome has steadily declined. It is likely that more and more individuals will wish either to select for or against the expression of a particular gene in their progeny. Through genetic counseling and testing and by sharing genotypes with potential mates, couples could either increase or decrease the potential for giving birth to a child with a propensity for a particular gene.

Thus, two persons, each carriers of a recessive gene for deafness, could enhance their percentages of having a deaf child by choosing one another as biological partners with the intent of having a deaf baby.

Genetic information can also play an important role in choices made by potential parents who choose In Vitro fertilization (IVF) as a method of obtaining a pregnancy. Ethical controversy continues to swirl as to whether parents who are deaf should be able to utilize the services of reproductive biology and IVF centers to assist them in having a deaf baby. This goal could be achieved prior to implantation by aspirating one cell from each embryo created from ova and sperm of the biological couple, fertilized and incubated in the laboratory. Genetic testing can be performed on the cells collected from the embryos to determine if the embryo is positive for the trait of deafness, and only implanting affected embryos. A second method of achieving this goal would be to monitor any pregnancy of the deaf couple using prenatal genetic testing with the intent of terminating any pregnancy with a normal hearing genotype and preserving any pregnancy with a deaf genotype. Some IVF centers have refused on ethical grounds to provide such services claiming that by design, the goal is to give birth to a child with a major disability, deafness. Other centers claim they will respect the autonomy of the couple and work with them to help attain goals that are motivated by their strong sense of valuing a deaf child. Of course, a male and female who are both carriers of an autosomal-dominant gene for deafness could become pregnant through natural reproductive methods. Even as Alexander Graham Bell was unsuccessful in getting legislation to ban deaf marriages in his day, it is even more unlikely that public policies would be enacted to prevent such births in this time.

Cost is an issue that figures into the ethical considerations of cochlear implants. Now that there is evidence that cochlear implants provide the benefits of improved hearing of sounds and improved speech, common sense suggests that two would be better than one. Indeed, especially within the past decade, a number of patients with a single cochlear implant returned to the surgeon to have bilateral implants. NIH-supported scientists found that the benefits of the cochlear implant far outweigh its costs in children. “A cochlear implant costs approximately $60,000 (including the surgery, adjustments, and training). In comparison, the services, special education, and adaptation related to a child that is deaf before age 3 costs more than $1 million” (NIH 2010).

Culture is likely to remain an issue although to a lessening degree. Members of the Deaf community claim that cochlear implants undermine and jeopardize their very existence. Members of the Danish Deaf Association have claimed that “deaf children are not sick or weak children, but normal Danish children, who just happen to use another language” (Nunes 2001).

Deaf activists have argued that from the ethical principle of respect for persons, minority cultures should be preserved and respect should be shown for the minority language (ASL). Many are also concerned that if there are receding numbers of deaf individuals and a decline in deaf culture, there will be less incentive for the majority culture to provide services and assistance that allow the deaf to flourish within the larger culture – a matter of justice.

Conclusion

In this section on ethical issues, the importance of ethical principles as moral action guides in ethical decision-making has been emphasized. Respect for persons and, in particular, the autonomy or self-determination of persons with regard to their ability to freely make health care choices is highly prized in this society. Likewise, the freedom of parents to make health care decisions for their minor children is valued. However, parental decision-making for deaf children should be informed by scientific evidence concerning harms and benefits associated with the use of cochlear implants, as implied by the principles of nonmaleficence and beneficence. There is now sufficient evidence that early implantation in deaf children is clearly a benefit, assisting them in achieving both receptive and expressive communication similar to their age-related hearing peers. The principle of justice promotes a duty to treat persons and classes of persons fairly. There is lingering debate and controversy about the appropriateness of considering the Deaf community as a unique culture and whether deafness is a disability that ought to be treated. Just as the medical community is committed to providing treatment that improves vision for those with difficulty seeing, it is committed to improving the capacity to hear for persons who are deaf. The innovation of auditory prostheses in the service of deaf persons is relatively new. Researchers continue to search for inventive ways to serve this population. The new "hybrid" cochlear implant is but one example. Researchers continue to experiment with auditory brainstem implants (ABI), and more studies regarding efficacy and safety will be required in the future. Further studies are warranted, but scientific research has demonstrated both the efficacy and safety of cochlear implants when carried out by well-trained surgeons and it is predicted that bilateral transplants will become a standard of treatment in the future.

Cross-References

▶ Ethical Implications of Sensory Prostheses
▶ Ethical Issues in Auditory Prostheses

References

Anthem Blue Cross-Blue Shield Medical Policy. (2012, July 10). *Cochlear implants and auditory brainstem implants*. http://www.anthem.com/medicalpolicies/policies/mp_pw_a050199.htm. Accessed 12 Nov 2012.

Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. New York: Oxford University Press.

Clark, G. (1969). *Middle ear and neural mechanisms in hearing and in the management of deafness*. Doctor of Philosophy thesis, University of Sydney, Sidney, Australia.

Clark, G. (2003). *Cochlear implants: Fundamentals and applications*. New York: Springer Science & Business Media.

Crouch, A., & Greenwald, B. H. (2002). Hearing with the eye: The rise of deaf education in the United States. In J. V. Van Cleve (Ed.), *The deaf history reader*. Washington, DC: Gallaudet Press.

Eisen, M. D. (2006). History of the cochlear implant. In S. B. Waltzman & J. T. Roland (Eds.), *Cochlear implants*. New York: Thieme Medical Publishers.

Eveleth, R. (2012, July 24). Should Oscar Pistorius's Prosthetic Legs Disqualify Him from the Olympics? Scientific American.

FDA (2002, July 24). *Public health web notification: Cochlear implant recipients may be at greater risk for meningitis*. US Food and Drug Administration Center for Devices and Radiological Health, Originally Issues. Updated October 17, 2002, http://ww.tsbvi.edu/seehear/winter03/fda.htm. Accessed 11 Nov 2012.

Fitzpatrick, E. M., Jacques, J., & Neuss, D. (2011). Parental perspectives on decision-making and outcomes in pediatric bilateral cochlear implantation. *International Journal of Audiology, 50*, 679–687. doi:10.3109/14992027.2011.590823.

Gannon, J. R. (1981). *Deaf heritage*. Silver Spring: Publishers National Association of the Deaf.

Gantz, B. J., & Turner, C. (2004). Combining acoustic and electrical speech processing: Iowa/Nucleus hybrid implant. *Acta Otolaryngology, 124*, 344–347.

Graham-Rowe, D. (2004). *First brainstem implants aim to tackle deafness*, New Scientist.http://www.newscientist.com/article/dn4540-first-brainstem-implants. Accessed 03 Nov 2012.

Hansson, S. O. (2005). Implant ethics. *Journal of Medical Ethics, 31*, 519–525.

Lang, H. G. (2007). Genesis of a community: The American deaf experience in the seventeenth and eighteenth centuries. In J. V. Van Cleve (Ed.), *The deaf history reader*. Washington, DC: Gallaudet University Press.

Lantos, J. D. (2012). Ethics for the pediatrician: The evolving ethics of cochlear implants in children. *Pediatrics in Review, 33*(7), 323–326.

Lee, C. (2012). Deafness and cochlear implants: A deaf scholar's perspective. *Journal of Child Neurology*, 27(81), 821–823. http://jcn.sagepub.com/content/27/6/821. Accessed 19 Oct 2012.

Lu, T., Djalilian, H., & Zeng, F. G. (2011). *An integrated vestibular-cochlear prosthesis for restoring balance and hearing*. 33rd annual international conference of the IEEE EMBS, Boston, MA.

McCormick, T. R. (2010). Ethical conflicts in caring for patients with cochlear implants. *Journal of Otology and Neurotology, 31*(8), 1184–1189.

Melton, M. F., & Backous, D. D. (2011). Preventing complications in pediatric cochlear implantation. *Current Opinion in Otolaryngology & Head and Neck Surgery, 19*, 358–362.

Merzenich, M. M., Michelson, R. P., Pettit, C. R., Schindler, R. A., & Reid, M. (1973). Neural encoding of sound sensation evoked by electrical stimulation of the acoustic nerve. *The Annals of Otology, Rhinology, and Laryngology, 822*, 486–503.

NAD Cochlear Implant Committee. (2000). http://www.nad.org/issues/technology/assistive-listening/cochlear-implants. Accessed 15 Oct 2012.

NAD. (2000). *Position statement on cochlear implants*. Published on National Association of the Deaf. http://www.nad.org. Accessed 15 Oct 2012.

NIH. (2010). *Yesterday, today & tomorrow*. U.S. Department of Health and Human Services. http://www.report.nih.gov/nihfactsheets/ViewFactSheet.aspx?csid=83. Accessed 20 Oct 2012.

Nunes, R. (2001). Ethical dimension of poediatric cochlear implantation. *Theoretical Medicine, 22*, 337–349.

Nuzzo, R. (2010). Cochlear implants can restore hearing. In C. F. Naff (Ed.), *Perspectives on diseases and disorders* (p. 54). Farmington Hills, MI: Gale Centage Learning/Greenhaven Press.

Punch, R., & Hyde, M. B. (2011). Communication, psychosocial, and educational outcomes for children with cochlear implants and challenges remaining for professionals and parents. *International Journal of Otolaryngology, 2011*, 1–10.

Salas, H. S. (2011). Cochlear implants and deaf children in clinical ethics. In D. S. Diekema, M. R. Mercurio, & M. B. Adam (Eds.), *Pediatrics: A case-based textbook*. Cambridge: Cambridge University Press.

US Dept of Health and Human Services, National Institutes of Health, Cochlear Implants. NIDCD Fact Sheet, NIH Publication No 11–4798 Updated November 2013 http://www.nidcd.nih.gov/health/hearing/pages/coch.aspx.

Waltzman, S. B. (2006). Speech perception in children with cochlear implants. In S. B. Waltzman & J. T. Roland (Eds.), *Cochlear implants*. New York: Thieme Medical Publishers.

Waltzman, S. B., & Roland, J. T. (Eds.). (2006). *Cochlear implants* (2nd ed.). New York: Thieme Medical Publishers.

Ethical Issues in Cochlear Implantation 50

Linda Komesaroff, Paul A. Komesaroff, and Merv Hyde

Contents

Abstract

Conventional perspectives regarding the ethical issues around cochlear implantation are often simplistic and under-informed. A common assumption is that there is a choice between unequivocal acceptance of the benefit of the technology on the one hand, and a choice based on respect for the integrity and values of the Deaf community on the other. The parameters of perceived choice are explored and it is concluded that such a simplistic model fails to recognize the

L. Komesaroff (✉)
Deakin University, Waurn Ponds, VIC, Australia
e-mail: wildpatch@gmail.com

P.A. Komesaroff
Monash Centre for Ethics in Medicine and Society, Monash University, Caulfield East, VIC, Australia
e-mail: paul.komesaroff@monash.edu

M. Hyde
Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Sippy Downs, QLD, Australia
e-mail: mhyde@usc.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_125,

complex and heterogeneous array of variables that need to be considered by, and the options that are available for, the individuals concerned, their families, and the medical practitioners advising them.

Introduction

The question of whether it is ethically appropriate to undertake cochlear implantation (see ▶ Chap. 49, "Ethical Issues in Auditory Prostheses", this volume) in young deaf babies has over a number of years been the subject of intense debate, and surprisingly little consensus has emerged. On the one side, those representing or supporting the medical community argue that deafness is a pathology, and that being able to hear is clearly *better* than being deaf, and that any therapy or technology which can help *cure* it should therefore surely be favored. On the other side, members of the Deaf community argue that being deaf is a legitimate way to be, that the meanings and values available to them, and which they share within well-functioning communities with high levels of support and solidarity, are as valid and valuable as those available to the hearing majority, and that technologies that are imposed on deaf people therefore fail to acknowledge and respect the validity and integrity of their distinct cultural resources.

Which side is right? Do the parents of a newly diagnosed deaf baby have a moral obligation to ensure that their child receives an implant at the earliest possible opportunity? Or should they rather exercise their (or their child's) cultural and linguistic right to decline the intervention and the technocratic disciplines and definition of identity that comes with it? Alternatively, should deaf children, as a matter of course, be provided with the possibility of enjoying the benefits of hearing, both in relation to their social functioning and their sensory experience? Indeed, do parents have a moral responsibility actively to support implantation (Tucker 1998)? Or is it more complex, with deaf children and their parents having more choices and alternatives than this simple dichotomy would suggest?

In public discussions about cochlear implantation, the views of hearing communities have enjoyed a natural, and obvious, advantage. The spectacular nature of the technology and the fascination with science, the public adulation of the scientist/businessmen who have promoted the technology, the access of eloquent, articulate spokespersons to the media, and humanistic discourses around curing sick children and providing expansive new opportunities for the disadvantaged, have – not surprisingly – struck a chord with both government and the public. The issue has even proved a convenient and effective device for children's hospitals – always strapped for money – to employ as part of their fund-raising campaigns.

With fewer resources at their disposal, members of the Deaf community and the organizations that represent them have done their best to challenge the cultural and linguistic assumptions underlying these positions. Their case against childhood implantation points to the validity of being a signing member of the hearing community and calls for parents to be provided with information about the Deaf community at the time of diagnosis (or discovery) of their child's deafness.

They have sought an emphasis on access to education and social services through their native sign language rather than on medical interventions that are directed toward enhancing their functioning within dominant (hearing/speaking) cultures. Attempts to move the argument beyond technical, medical, and audiological issues highlight the importance of deaf culture and the central place of a native sign language in deaf people's lives (see Lane and Bahan 1998; Ladd 2007). A recent resolution of the World Federation of the Deaf (2011) reaffirmed the need for information on sign language development to be provided to parents of children with cochlear implants, drawing on a claim to their "rights" as members of a cultural and linguistic minority. This perspective is supported by several conventions of the United Nations (see Komesaroff 2000), in particular the UN Convention on the Rights of People with a Disability (United Nations 2006; see Hyde 2007), which asserts the rights of sign language users to access and participate in the full range of life experiences. Article 24, section 3(c) of the convention calls for the education of children who are deaf to be "delivered in the most appropriate languages and modes and means of communication for the individual, and in environments which maximize academic and social development."

There can be no doubt that the pathologization of deafness has reinforced the assumption (unsupported by research; see Hyde 2005) that participation in a signing culture will undermine the ability of deaf children to acquire and use speech and, as a result, to function and thrive in (hearing) society. As such, a common myth is that children with implants need to be educated in a non-signing environment to avoid interference from sign language on spoken language development that has resulted in an "endangerment" of sign languages (see Murray n.d.). Cochlear implantation is the dominant approach to treating congenital deafness among people in most Western countries (Komesaroff 2007b) and while the counter-arguments of the Deaf community that deafness should be regarded as an example of cultural difference rather than as a medical pathology have drawn attention to the existence of a contrary perspective, these appear to have done little to disturb or disrupt the status quo and majority cultural assumptions have prevailed.

As a result, the dilemmas faced by parents and medical practitioners remain as stark as ever. The problem is posed as one requiring a clear decision between two fundamentally opposed cultural and ethical positions. This has led to considerable uncertainty and anxiety among all concerned, but especially among the parents of newly diagnosed deaf babies or children. How should they respond to this dilemma of pathology or culture, on which the future welfare and happiness of their children would seem to depend? How should they manage such ambiguity? How do they – and we – answer when asked "What should we do?."

As with other apparently intractable ethical dilemmas, in this one, there is a need to consider the assumptions lying behind the issues and to explore the possibilities for dialogue between the two opposing positions. We need to reflect on the frameworks of knowledge and culture that have generated the two viewpoints and open up the possibility of a third way that incorporates respect for both. We need to be able to accept that implantation, not unlike other areas of medical intervention, can be problematic, but that so also is the unconditional rejection of implantation

(see ▶ Chap. 48, "Ethical Implications of Sensory Prostheses", this volume). In the process, we are compelled to disrupt both the conventional medical view of deafness as pathology (with its associated social controls implicit in surgical intervention) and the cultural view of deaf people as members of an autochthonous, independently functioning, linguistic, and cultural minority. Dialogue between the two discourses will require recognition of the possibility of complex, multifaceted identities that enable individuals to inhabit multiple social groups simultaneously and to operate competently in diverse cultural domains.

In reality, neither the Hearing nor the Deaf World is unitary or homogenous. Cochlear implantation exists within diverse, complex, and changing social environments in which new meanings are constantly being negotiated and contested. The processes by which such meanings arise and pass away, furthermore – as in all negotiations around ethical values – are highly dependent on local conditions, including individual needs, personal preferences, and cultural biases. These can be unpredictable and subject to change, along with the specific circumstances that apply to particular cases. If we are to develop an ethical discourse around cochlear implantation, therefore, that is capable of encompassing the full complexity both of the issues and of the decision-making processes associated with them, we need to go beyond the conventional formulations and the mere acknowledgment of linguistic and cultural rights. Instead, we will be obliged to question prevailing assumptions about deafness, on all sides, and to find ways to actively promote the deep cultural and ethical dialogues by which difference is realized and individuality developed and expressed.

What is the nature of a deaf life? How can and do deaf people contribute to the lives of all people? Can cochlear implants be viewed as one of a multitude of devices or tools to be used at the discretion of deaf people, depending on their purposes, social contexts, and value frameworks? Deconstruction of existing value systems allows us to expand the possibilities for ethical action (Komesaroff 2008). This does not lead to definitive, outcome-focused resolutions of ethical dilemmas but to the reactivation of a process within which productive communication and dialogue across difference can occur. Such an approach emphasizes the multiple aspects of implantation itself along with the multiple readings of deafness, together with the sequence of decisions that must be made by parents and deaf people (as against a unitary "decision"). Opposite and contradictory points of view necessarily abound – and they need to coexist. Examination of the lived experience of deaf people's lives allows reductionist views of "deafness versus hearing" to be challenged and new directions for the medical practice of implantation to be opened up.

Historical Perspective

Aside from the "normal" risks and insured liabilities associated with the conduct of any surgery, the introduction of ethical considerations into cochlear implantation was initially seen as unnecessary or even intrusive

(Power and Hyde 1992; Hyde 1995). Medical clinicians, parents of young deaf children, and even some deaf individuals themselves, all saw the debates about ethics as imposing an unwanted limitation on their right to choose. After all, the process and its possible outcomes seemed obviously to be "good" (Hyde and Power 2000).

At least in this regard, 20 years later, a lot has changed. Against this initially hostile context, considerable progress has been made by all the key stakeholders. The claims of "genocide" (see Lane 2005; Ladd 2007), asserted in the heat of the divisive and emotive rhetoric by some groups against surgeons and implant programs, have been shown to be extravagant and unfounded. At the same time, the depiction of implantation as the "cure to deafness," so commonly portrayed in the media (Komesaroff 2007), has also proven to be incorrect and exaggerated. An elaborated understanding of both "risk" and "benefit" has been incorporated into the protocols used before surgery and into the consent procedures routinely observed with parents and deaf people (Punch and Hyde 2011a, b). The extent of the risk-benefit analysis has been significantly broadened to include recognition of the social, cultural, and educational dimensions along with the anatomical and surgical factors.

Nonetheless, there remains much to be done. Wider issues such as identity, well-being, and social inclusion and isolation need further exploration. The implications of research that demonstrates adverse educational and social consequences of a child's deafness, regardless of implantation, need to be considered. The recognition that deaf children with implants may achieve educational outcomes at the level of those associated with "hard of hearing" students (Hyde and Punch 2010) suggests that the analysis of risk and benefit needs to be extended still further, across a broader time frame, to allow the assessment of educational, social, and vocational outcomes into young adulthood (18–24 years of age).

An accurate understanding of the available facts is important. It has been argued by Deaf communities that the decision about whether or not to undergo a cochlear implant can be left until the individual concerned is an adult. In addition, medical professionals have advised parents that the learning of sign language can be delayed until later in life, according to preference. However, there are time constraints to both decisions. It is now generally accepted that surgery and rehabilitation must occur as early as possible in a deaf child's life if the best outcomes in the acquisition of a spoken language are to be attained. Similarly, there is a critical period within which a first language needs to be acquired (whether it is a spoken or signed language) for native-like fluency to be possible.

Rather than retaining an exclusive focus on whether or not to implant a young child, there is a growing recognition of the need for ongoing *re*-evaluation of the *multiple* decisions to be made throughout a deaf person's life. The initial decision about whether or not to implant is made by parents based on the information available to them at the time. As their deaf child grows and shows preferences for various means of communication, thrives or struggles with education, and identifies with some individuals or groups over others, he or

she may make unexpected and changing decisions about the use of their implant and how it relates to their developing identity and well-being (see ▶ Chap. 22, "Neuroethics and Identity", this volume).

In this context, it may be better and more appropriate to move away from the understanding of cochlear implants as a cure for deafness, and to regard them instead as an example of one of multiple technologies that can be used by deaf people to navigate their way through life. We now recognize that deaf people make use of implants in different ways at different times and in different contexts. Alongside their use of an implant, they may use sign language, interpreters, computers, telecommunication devices, and so on – depending on what is needed and what works best at the time, and in relation to particular social, recreational, and vocational needs. The ability to move seamlessly between different means of communication is an important indicator of both the acquisition of effective competence as a communicator and of mature and robust personal identity.

The Theory of Minimal Adaptation, proposed by Bernard Farber and retained in use by more recent researchers (Farber 1960; Brown et al. 2008), may provide a useful framework for explaining the ways in which people from differing perspectives can respond to challenges in constructive and nonthreatening ways. According to this theory, individuals typically try to solve issues in incremental steps, beginning with changes that differ in the least possible way from their current positions. For example, to hearing parents, providing their child with a cochlear implant may initially offer the least adaptation necessary to preserve *their* cultural and linguistic heritage and aspirations for the child. As time goes by, and further changes typically become necessary, however, they may find it possible to shift more definitively from their initial position or decision to embrace sign language and deaf culture, ultimately moving further toward what is considered a more "Deaf World" view.

If this is how decision-making occurs in practice – and indeed, is encouraged – the time frames in which the risk-benefit analysis around cochlear implantation is conceived and the consent processes are negotiated need to be extended. As deaf infants with cochlear implants grow and develop, their identities will unfold in relation to the range of experiences to which they are exposed and the opportunities they encounter. The development of a mature identity cannot be understood simplistically in terms of a choice between the identity of a "hearing person" and that of a "deaf person." Even expanded descriptions, (e.g., Bat-Chava 2000), are inadequate and inauthentic and thus unable to reflect the more multilayered and complex outcomes that are likely.

There are many other factors that also need to be taken into consideration. For example, recent studies (Punch and Hyde 2011a, b) have identified a range of challenges that arise across the years of primary schooling, through adolescence and into emerging adulthood for deaf people with implants. This research suggests that issues of social isolation, some areas of school achievement (e.g., literacy and numeracy), and reduced social and emotional well-being may also need to be included in the discourses about ethics and shared meaning around implantation.

Emerging Adulthood, Resilience, and Well-Being

Of itself, hearing loss does not inevitably pose a risk to successful transition as an emerging adult. However, significant hearing loss sustained in childhood or adolescence creates a vulnerability that can limit the capacity of individuals to adapt to the demands and opportunities presented in emerging adulthood and threaten psychological well-being and achievements. Factors contributing to this vulnerability can include communication difficulties with family members and peers, a limited range of social relationships, feelings of isolation, reduced school achievement, lowered expectations by others, and feelings of dependency (Valentine and Skelton 2007). These may limit the development of "resilience" and contribute to stress, anxiety, and lowered measures of well-being and life satisfaction (Gascon-Ramos 2008). Deaf and hard of hearing (DHH) individuals within a hearing environment are invariably faced with daily obstacles and challenges relating to communication difficulties and a range of social and structural barriers, including stigma and prejudice, contributing to personal and social problems in emerging adulthood (Lukomski 2011; Meyer and Kashubeck-West 2011; Punch et al. 2004).

The mechanisms underlying the development of resilience are not entirely clear. Meyer and Kashubeck-West (2011) concluded that interactions between environmental and individual factors are of particular concern for the development of resilience in young adults who are deaf. However, others such as Valentine and Skelton (2007) and Gascon-Ramos (2008) view resilience and well-being as a consequence or outcome of the identity that a DHH child develops; in this view, the use of a sign language and exposure to a deaf culture is seen as being causally related to a stronger sense of identity and subsequently greater resilience and psychological well-being through emerging adulthood. In contrast, Hyde and Power (2006) argued that resilience can also be typified as the capacity to "overcome deafness" and to succeed, "despite" the presence of hearing loss. The assumption – they claimed – is often that resilience is associated with success in terms of the individual developing proficiency in the use of a spoken language and in circumstances where normally hearing adults are seen to succeed. Although this last vision of resilience is naturally appealing to many parents, it does not account for the diversity among people with hearing loss and the differing environments in which they develop and learn; furthermore, it can also be a strong stigmatizing influence for those who cannot reach anticipated standards and it can create lowered perceptions of self-efficacy and worth (Hyde and Power 2006; Punch and Hyde 2005).

Another area in which data are lacking is that of DHH emerging adults and their psychological well-being. Meyer and Kashubeck-West (2011) reviewed the small amount of research on emerging adults who are deaf and found that it focused on pathological indicators such as might be observed in psychological distress or during periods of unemployment. They found that very little emphasis has been placed on potentially positive psychological constructs such as the development of well-being and life satisfaction.

Emerging Adulthood and Identity

In many countries, the great majority of DHH children grow up in homes and go to schools with normally hearing members, without exposure to a sign language and with the expectation by both parents and teachers that a "hearing" identity is the objective or desirable outcome (Hyde and Power 2003; Punch and Hyde 2011a). The almost universal use of cochlear implants in significantly deaf infants and young children has further entrenched this expectation (Hyde et al. 2010; Punch and Hyde 2010). Such an implicit and unquestioned assumption of a key ethical goal, however, often sets the scene for a clash with the demand for greater personal autonomy associated with adolescence, which for young DHH people involves attempts to adapt to the transitions and challenges associated with emerging adulthood within expanding social, vocational, and educational networks, including engagement with other young people facing similar issues.

In this context, some studies suggest that the use of a sign language and identification with a Deaf community can be associated with higher levels of self-esteem and psychological well-being (Valentine and Skelton 2007). Other researchers have found four specific deaf identities (Fischer and McWhirter 2001; Glickman and Carey 1993): "hearing," "marginal," "immersion," and "bicultural" – to describe individuals' "orientation[s] to an affiliation with the Deaf Community and Deaf Culture" (Fischer et al. 2001, p. 358). In a similar vein, Bat-Chava (2000) described three identities: "culturally deaf," "culturally hearing," and "bicultural."

While these categorical models may offer some useful – albeit perhaps oversimplified – descriptions of the identities that young deaf people may develop, they do not take account of the complexity of the experiences of emerging deaf adults in contemporary societies. For example, a recent study (Hyde and Punch 2011) found that some hearing parents of children with cochlear implants valued the use of signing as a way for their children to establish a connection to other deaf people and, perhaps, a sense of belonging to the Deaf community, and that in adolescence, some of the young people with cochlear implants themselves for the first time developed an interest in the Deaf community.

Any simple description of the nature of the identity formation of emerging deaf adults would seem insufficient to describe the complex interplay of variables and opportunities experienced by young deaf people in contemporary post-school transitions. It is possible that other theories of identity may prove more applicable, such as that proposed by Cass (1996) to describe the isolation, confusion, and stigma often experienced by individuals with alternative sexual orientations, or the intersectionality theory of Crenshaw (1991), which suggests that various biological, social, and cultural categories do not act alone and may interact in different environments to form more complex accommodations of identity.

Cochlear Implantation, Transition, and Well-Being

As we have seen, there are strong expectations among many parents, implant providers, intervention and educational service providers, and some deaf people themselves that the use of cochlear implants will "conquer deafness" (Wheeler et al. 2009). Research findings have shown significant benefits on many measures, but not as yet universal outcomes from the perspectives of parents, teachers, or young deaf people themselves (see Punch and Hyde 2011b for a review). A recent Australian study has found that social participation and emotional well-being remain consistently problematic as children with cochlear implants reach adolescence. It appears that even cochlear implant users with excellent spoken language development experience the phenomenon of "social deafness," a term describing the effects of hearing loss in social interactions involving groups of people or in noisy environments, such as in post-school education, work and social environments (Punch and Hyde 2011a). Although potential school-to-work transition, career development, and psychological well-being outcomes may differ among DHH young people with and without implants, many issues, challenges, and barriers are likely to be common to both groups.

In particular, DHH young adults are at greater risk of attrition from further education and training, of unemployment and underemployment, and of lower levels of psychological well-being and life satisfaction. Further, there has been limited study of the period of emerging adulthood as experienced by DHH people who have been sustained users of cochlear implants. To realize the benefit of major, ongoing public investment in health and education systems, ameliorate the effects of growing up with hearing loss and achieve economic parity, further data are needed in these areas before a more complete risk-benefit analysis can be undertaken and incorporated into consent processes.

Conclusion

We have argued that conventional formulations regarding the ethical issues around cochlear implantation are excessively simplistic. The assumption that in every case, a simple and unequivocal choice has to be made between unconditional acceptance of the technology and the hegemony of the hearing culture, on the one hand, and respect for the integrity and values of the Deaf community, on the other, fails to recognize the complex and heterogeneous array of variables that need to be considered by, and the choices that are available for, the individuals concerned, their families, and the medical practitioners advising them. The familiar discussions focusing on abstract, context-independent categories such as autonomy, informed consent, and beneficence, furthermore, systematically obscure the uncertainties of knowledge and science, the variety of value positions, the ambivalence commonly experienced by parents and carers, and the varying psychological and cultural needs of deaf and hard of hearing children and their families.

The process of dialogue around ethical issues is complex, and involves many factors, such as individual circumstances, value preferences, and social and cultural settings, which often evolve and change over time. Unsurprisingly, it is rarely the case that there is a single point of view that is clearly preferable to all others; rather, each typically contains elements of truth as well as tensions and uncertainties that need to be resolved in relation to the prevailing circumstances and the available resources. Through this dense plenum of values that characterizes the ethical territory of hearing and deafness, individuals and families negotiate their trajectories. In most cases, they do not find themselves compelled to make an irrevocable commitment to a single, preformed identity: between, for example, that of a person who accepts without question the benefits of technology or that of membership of a signing community. On the contrary, for the most part, identities are composed and re-composed according to circumstances, experiences, and changing conditions; they may be fluid and multifaceted; and they may remain subject to uncertainty, ambivalence, and critical questioning.

Typically, decisions are made locally, in relation to immediate circumstances and challenges, rather than in reference to large-scale principles or long-term anticipated outcomes. Accordingly, the ethical reflections that support these decisions are not purely theoretical. They depend sensitively on local conditions and personal preferences, as well as on knowledge and the cultural valorization of alternative courses of action. Detailed knowledge of the findings of empirical research into the various factors influencing the welfare of young people and developing adults is essential for an adequate understanding both of the issues themselves and of effective responses to them, including research into the conditions for the emergence of resilience and psychosocial well-being, and problems and challenges in relation to education. In other words, ethical discourse involves a relentless process of exposing questions, of clarifying the empirical conditions that shape the possible outcomes, and the subsequent mapping of the ethical and social meanings thereby generated.

The time has come to move on from the rigid polarization between implantation and Deaf cultural characteristics such as the use of sign languages to more open, tolerant flexible arrangements that incorporate hybrid combinations of the various possibilities within a cultural field that supports and stimulates ongoing critical reflection and review.

Cross-References

► Ethical Implications of Sensory Prostheses
► Ethical Issues in Auditory Prostheses
► Neuroethics and Identity

References

Bat-Chava, Y. (2000). Diversity of deaf identities. *American Annals of the Deaf, 145*, 420–428.

Brown, W. H., Odom, S. L., & McConnell, S. R. (Eds.). (2008). *Social competence of young children: Risk, disability, and intervention*. Baltimore: Brookes.

Cass, V. (1996). Sexual orientation identity formation: A Western phenomenon. In R. P. Caba & T. S. Stein (Eds.), *Textbook of homosexuality and mental health* (pp. 227–251). Washington, DC: American Psychiatric Association.

Crenshaw, K. W. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241–1299.

Farber, B. (1960). Effects of a severely mentally retarded child on family integration. *Monographs of the Society for Research in Child Development, 24*, 1–112.

Fischer, L. C., & McWhirter, J. J. (2001). The deaf identity development scale: A revision and validation. *Journal of Counseling Psychology, 48*, 355–358.

Gascon-Ramos, M. (2008). Wellbeing in deaf children: A framework of understanding. *Education and Child Psychology, 25*(2), 57–71.

Glickman, N. S., & Carey, J. C. (1993). Measuring deaf cultural identities: A preliminary investigation. *Rehabilitation Psychology, 38*(4), 275–283.

Hyde, M. B. (1995). Ethical dimensions of cochlear implantation of deaf children. *Annals of Otology, Rhinology and Laryngology – Supplement, 104*, 19–20.

Hyde, M. B. (2005). *Is sign language incompatible with the effective use of a CI?* Nordic conference on children's emotional and intellectual development through language. Torshavn, Faroe Islands, August 29–30, 2005.

Hyde, M. B. (2007). *Deafness and human rights. How the 2007 UN convention on the rights of persons with disabilities may influence current policies and programs*. Keynote presentation at the Nordic Conference: The dual languages of deaf and hearing-impaired children. Gothenburg, Sweden, September 3, 2007.

Hyde, M. B., & Power, D. J. (2000). Informed parental consent for cochlear implantation of deaf children. *Australian Journal of Social Issues, 35*(2), 117–128.

Hyde, M., & Power, D. (2003). Characteristics of deaf and hard of hearing students in Australian regular schools: Hearing level comparisons. *Deafness and Education International, 5*, 133–143.

Hyde, M., & Power, D. (2006). Some ethical dimensions of cochlear implantation for deaf children and their families. *Journal of Deaf Studies and Deaf Education, 11*, 102–111.

Hyde, M., & Punch, R. (2010). Children with cochlear implants in Australia: Educational settings, supports, and outcomes. *Journal of Deaf Studies and Deaf Education, 15*(4), 405–421.

Hyde, M. B., & Punch, R. (2011). The modes of communication used by children with cochlear implants and role of sign in their lives. *American Annals of the Deaf, 155*(5), 535–549.

Hyde, M., Punch, R., & Komesaroff, L. (2010). A comparison of the anticipated benefits and received outcomes of paediatric cochlear implantation: Parental perspectives. *American Annals of the Deaf, 155*(3), 322–338.

Hyde, M., Punch, R., & Komesaroff, L. (2010). Coming to a decision about cochlear implantation: Parents making choices for their deaf children. *Journal of Deaf Studies and Deaf Education, 15*(2), 162–178.

Komesaroff, L. (2000). Linguistic rights of the deaf: Struggling against disabling pedagogy in education. *Australian Journal of Human Rights, 6*(1), 59–78.

Komesaroff, L. (2007a). Media representation and cochlear implantation. In L. Komesaroff (Ed.), *Surgical consent: Bioethics and cochlear implantation* (pp. 88–119). Washington, DC: Gallaudet Univerity Press.

Komesaroff, L. (Ed.). (2007b). *Surgical consent: Bioethics and cochlear implantation*. Washington, DC: Gallaudet Univerity Press.

Komesaroff, P. A. (2008). *Experiments in love and death: Medicine, postmodernism, microethics and the body*. Carlton: Melbourne University Press.

Ladd, P. (2007). Cochlear implantation, colonialism, and deaf rights. In L. Komesaroff (Ed.), *Surgical consent: Bioethics and cochlear implantation* (pp. 1–29). Washington, DC: Gallaudet Univerity Press.

Lane, H. (2005). Ethnicity, ethics, and the Deaf-World. *Journal of Deaf Studies and Deaf Education, 10*(3), 291–310.

Lane, H., & Bahan, B. (1998). Ethics of cochlear implantation in young children: A review and reply from a Deaf-World perspective. *Otolaryngology – Head and Neck Surgery, 119*, 297–313.

Lukomski, J. (2011). Resiliency and the emerging deaf adult. In D. H. Zand & K. J. Pierce (Eds.), *Resilience in deaf children* (pp. 375–390). New York: Springer.

Meyer, J., & Kashubeck-West, S. (2011). Psychological wellbeing in emerging adults who are deaf. In D. H. Zand & K. J. Pierce (Eds.), *Resilience in deaf children* (pp. 359–374). New York: Springer.

Murray, J. J. (n.d.). *Conference summary: Sign languages as endangered languages*. Resource document. World Federation of the Deaf. http://wfdeaf.org/news/conference-summary-sign-languages-as-endangered-languages. Accessed 10 Jan 2013.

Power, D. J., & Hyde, M. B. (1992). The deaf community and cochlear implants. *Medical Journal of Australia, 157*, 421–422.

Punch, R., & Hyde, M. (2005). The social participation and career decision-making of hard-of-hearing adolescents in regular classes. *Deafness and Education International, 7*(3), 122–138.

Punch, R., & Hyde, M. (2011a). Communication, psychosocial and educational outcomes of children with cochlear implants and challenges remaining for professionals and parents. *International Journal of Otolaryngology*, Article ID 573280, 10.

Punch, R., & Hyde, M. (2011b). Social participation of children with cochlear implants: A qualitative analysis of parent, teacher, and child interviews. *Journal of Deaf Studies and Deaf Education, 16*(4), 474–493.

Punch, R., Hyde, M. B., & Creed, P. (2004). Issues in the school-to-work transition of hard-of-hearing adolescents. *American Annals of the Deaf, 149*, 1.

Tucker, B. (1998). Deaf culture, cochlear implants and elective disability. *Hastings Center Report, 28*(4), 6–14.

United Nations. (2006). *Convention on the rights of people with disabilities*. Resource Document. United Nations. http://www.un.org/disabilities/documents/convention/convoptprot-e.pdf. Accessed 10 May 2011.

Valentine, G., & Skelton, T. (2007). Re-defining norms: D/deaf young people's transitions to independence. *The Sociological Review, 55*, 104–123.

Wheeler, A., Archbold, S. M., Hardie, T., & Watson, L. M. (2009). Children with cochlear implants: The communication journey. *Cochlear Implants International, 10*(1), 41–62.

World Federation of the Deaf. (2011). *Congress Resolution: XVI World Congress of the World Federation of the Deaf*. Resource document. World Federation of the Deaf. http://wfdeaf.org/news/congress-resolution. Accessed 10 Jan 2013.

Sensory Enhancement

51

Karim Jebari

Contents

Abstract

Sensory enhancement is a form of human enhancement that aims to extend the sensory capabilities of a person beyond what is possible for a normal human. Sensory enhancement can consist in either an enhancement that improves a sense or that extends that sense to perceive light, sound, tactile stimuli, or chemical traces that are beyond the human range. A sensory enhancement may also add a new sense, such as electroreception or modulate a sense so that it can perform completely new functions, such as echolocation (biosonar). This chapter argues that sensory enhancement could be implemented in mainly two ways: either via the application of digital technology or by genetic engineering of the human body. The potential of augmented reality (AR) and of brain-computer Interface (BCI) technology is also explored in the section on digital

K. Jebari
Department of Philosophy, Royal Institute of Technology (KTH), Stockholm, Sweden
e-mail: Jebari@kth.se; jebarikarim@gmail.com; karim.jebari@abe.kth.se

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_106,

enhancement. The section on genetic engineering will mainly be concerned with the potential of horizontal gene transfer (HGT). Three arguments on the normative aspects of sensory enhancement are also presented in this chapter. The first considers the instrumental value of being able to perceive new forms of artistic expression. The second concerns the idea of diversity and whether sensory enhancement could increase human diversity. The third argument departs from the "capabilities approach," formulated by Amartya Sen and Martha Nussbaum, and sketches out the position that we may be deprived in comparison to some possible future enhanced people, even if we do not regret being so.

Introduction

Sensory enhancement is in this context an instance of *human enhancement*. Thus, a correction of deficient hearing or sight is not covered here. At the moment, glasses and cochlea implants are not sensory enhancement technologies. Should telescopes and microscopes be defined as sensory enhancement technologies? Here the distinction is more tenuous and based on what we perceive to be part of a person's abilities. For something to count as an enhancement, this must add or extend a person's functioning, or ability to do something. Clearly we do not believe that a person who owns a telescope has the ability to see the Galilean moons, even if that person is in some more general sense able to do it. Human enhancement technologies ought to therefore be seen as artifacts that are to a certain degree integrated with our persona. According to a narrow view, only things that are physically integrated, i.e., attached or assimilated to the body, ought to count. A clear example of this would be a vaccine. On a more inclusive view, devices that form part of a person's mental self-representation should also be included. On this view, if a person perceives glasses, clothes, contact lenses, and (perhaps) a smartphone to be part of his or her body, these devices count as human enhancements. This vague characterization will vary in different contexts and across generations, but it will suffice for the purposes of this discussion. Another distinction that ought to be considered is between a sensory enhancement that allows us to perceive a sensory input without giving us a new sensory quality and an enhancement that would literally change how we see the world. For example, a thermographic camera forms an image of infrared light using visible light. An observer can therefore indirectly see infrared light. This sensory experience probably differs from what it might be like to be able to perceive infrared directly, which of course may differ radically between different individuals and species. However, an indirect experience of the world may be just as useful for practical purposes as a direct one. We should therefore include technologies that allow people to "see" indirectly in our definition of sensory enhancement. Another relevant distinction is between enhancement of our *sensory capacities* and the enhancement of our *perception*. Whereas the first changes *what* we perceive, the second changes *how* we perceive. As an example of perceptual enhancement,

consider some drugs that allegedly enhance the way we perceive music, and allow us to distinguish between subtle nuances. The difference here is that whereas sensory enhancement is primarily informational, perceptual enhancement is primarily phenomenological. This chapter will not be concerned with perceptual enhancement.

Some of the proposals that have been made for human enhancement refer to new and improved sensory functions. For example, concrete research is currently being performed on a bionic contact lens the display of which will be superimposed on what one sees naturally. Other proposals include improved vision (ultraviolet and/or infrared wavelengths), more acute hearing, and chemical sensors. Although such possibilities have been mentioned in the literature on enhancement, sensory enhancement is not the main topic of a single publication that is tracked in the philosophical database *phil papers* or the medical database *pub med*. This chapter will argue that this lack of attention is regrettable, due to the potential importance of sensory enhancement. It will isolate the issue of sensory enhancement from the more general issue of enhancement, covering both the more speculative and the more realistic possibilities for enhancement, but with a clear emphasis on the latter.

Outline of the Opportunity Space

Any plausible account of future developments demarcates the possible and interesting from the purely fantastical. This is no trivial task. Yet, any discussion of this nature requires that we anchor our expectations along some space of possible outcomes. I suggest a possible heuristic for doing this in the context of sensory enhancement. Although physics sets a definitive boundary from what we can expect to be possible, the boundary set by biology is narrower and more interesting. More specifically, I propose that when thinking about the possibility space of potential sensory enhancements, we confine ourselves by the boundary set by animals in the phylum *chordate* (vertebrates). While the species diversity within this phylum is formidable, ranging from fish and bird to mammals, these animals share a relatively well-developed brain, complex sensory organs, and a circulatory system with a heart (Romer and Parsons 1986). According to the proposed "chordate-heuristic," a biological function in the realm of sensory perception that is possible in this phylum might be possible for humans in the future, given the appropriate technology.

In this chapter, two technology trends are extrapolated to provide a plausible basis for these considerations: the rapidly advancing prowess of genetic engineering, in particular horizontal gene transfer, or HGT. This technology consists in the transfer of genes between organisms. The second technology trend is the miniaturization and price reduction of processing power and the probable future ubiquity of computers and sensors. Two technologies are of particular interest here. First, augmented reality, or AR, is computer-generated sensory input that is superimposed on reality. This makes mediation of sensory inputs possible.

The second technology is brain-computer interface, or BCI, which refers to a device that allows direct communication between the central nervous system and a computer by means of electrodes. Such devices are already used in cochlea implants and artificial retinas.

Digital Senses: From Augmented Reality to Brain-Computer Interface

Augmented Reality

As computers and digital sensors have become smaller, cheaper, and more powerful, their usefulness for enhancement purposes has become evident. It is reasonable to conjecture that any sensory data gathered by a machine that is either wearable or possible to connect to the internet could, with the appropriate interface, add to our own sensory experience. Increased digitalization has made sensory-enhancing technology cheaper and better. Night goggles are, for example, widely used by military forces and police officers all over the world. As these goggles and other similar technology increasingly rely on computers, we can reasonably expect that night vision might become cheaper and more suited for nonprofessional use. Digital hearing aids can enhance audition both by amplifying sounds and enhance signal to noise ratios in the environment. However, much more radical enhancement can be achieved by "outsourcing" our perceptual apparatus with the help of emerging technology. Consider the field of augmented reality or AR. Augmented reality is a live, direct or indirect, view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics, or GPS data. A head-mounted display (HMD) places images of both the physical world and registered virtual graphical objects over the user's view of the world. Project Glass is a research and development program by Google to develop an augmented reality head-mounted display (HMD). This product will, according to Google, be available in the consumer market in 2014 (Goldman 2012).

Bionic contact lenses are being developed to provide a virtual display that could have a variety of uses from assisting the visually impaired, to the video game industry. These devices will have the form of conventional contact lenses with added bionics technology. These lenses will eventually have functional electronic circuits and infrared lights to create a virtual display (Collier 2010). Leading researchers in this field state that "Looking through a completed lens, you would see what the display is generating superimposed on the world outside" (Hickey 2008). In 2011, Lingley et al. created and tested in vivo a functioning wirelessly powered prototype with a single-pixel display. These contact lenses were tested on rabbits and showed no adverse effects (Lingley et al. 2011).

The potential to enhance vision and hearing with the help of AR is quite significant. A likely development in the near future is what is sometimes referred to as "the internet of things" (Ashton 2009). The idea behind this concept is that

ordinary things will increasingly be equipped with identification tags, computers, and sensors that feed their information through the web. Thermal imaging cameras are already widely used as tools for surveillance, and are becoming so cheap that they are available as consumer products. Increasingly, these cameras are equipped with connectivity, allowing users to tap into them via Smartphone apps. Night vision cameras and other electronic equipment with sensors capable of obtaining sensory input beyond the human ability could easily be mounted on cars, streetlamps, and signposts and thus provide wireless information to AR systems, thereby providing users with night vision, vision in the UV-range, telescopic vision, and other kinds of enhancements. The ability to accurately measure distance, size, and determine the mass, density, and the trajectory of an object are typical examples of enhancements that are well suited to an AR system. Hearing could just as plausibly be outsourced to sensors embedded in the environment. Sensors could pick up ultra- and infrasound, enhance accuracy, and add information about the source of the sound.

Electric noses are devices that can effectively detect small concentration of chemicals in confined spaces. With direct applications in security, medicine, food production, and chemical safety, this technology has a clear commercial potential. With an appropriate interface, such as either AR or a brain-computer interface, this technology could be used to enhance our sense of smell. Like connecting an artificial limb with the somatosensory system, or directly stimulating ocular nerves to repair sight, olfactory enhancement has the potential to profoundly affect functioning. Whether or not these electric noses will enter the consumer market for enhancement purposes remains to be seen (Wilson and Baietto 2009).

Magnetoreception is a sense which allows an animal to detect a magnetic field to perceive direction, altitude, or location. It has been hypothesized that the ability to sense the Earth's magnetic field is essential to some birds' ability to navigate during migration (Walcott 1996). Magnetoreception in humans has been achieved by magnetic implants used as non-permanently attached artificial sensory "organs" (Nagel et al. 2005). Small neodymium magnets can be placed under the skin (usually the fingertips) so that the movement of the implant in the presence of magnetic fields can be felt by the individual. These implants can in this way be used to convert nonhuman sensory information into touch (Hameed et al. 2010). As with AR, this kind of enhancement does not provide the user with a new sense directly. Rather it allows a user to gather new information but via the user's sense of touch. These implants are to some extent used by subcultures that experiment with body modification (Norton 2006).

Brain-Computer Interface and Sensory Enhancement

Since AR superimposes sensory data in way that we can interpret, it only gives us a "superficial" enhancement. For example, a device that can allow a user to "see" UV light will represent this light through the user's AR device into light that the user is able to perceive. Although this might be interesting and useful, it does not

fundamentally alter our perceptual apparatus. However, brain-computer interface (BCI) based devices could do just that. A BCI connects the central nervous system directly to a computer, and allows information to be transferred between these two systems through electrical impulses. Cochlear implants are prosthetic devices that use this technology and that feed sensory information directly to the auditory nerve. BCI implants could in theory mediate sensory input without adapting it to the range of the human ear or eye. Thus, a BCI-mediated sensory experience would probably differ from that mediated via AR. Whereas we are likely to see AR enhancements of the kind discussed above in a few years, BCI enhancements require several scientific breakthroughs regarding both miniaturization and the understanding of the relevant mechanisms in the brain. In addition, since a BCI that provides this amount of information to specific parts of the brain requires an invasive procedure, we are not likely to see this kind of enhancement to be widely adopted unless a noninvasive (transcranial) method to accurately target parts in the brain associated with perception is devised. Although a novel interest for noninvasive electrical stimulation has had a kind of renaissance, neither of the two main technologies, i.e., transcranial direct-current stimulation and transcranial magnetic stimulation, have yet been used to produce sensory experiences similar to those of a cochlea implant.

Gene Transfer

Horizontal gene transfer (HGT) consists of the transfer of genetic material between organisms. Since genes are in sense instructions to produce proteins, new genes imply new proteins and thus new abilities. For example, this is the technique behind GloFish, a zebra fish with genes that encodes the green florescent protein, originally extracted from jellyfish (Zhiyuan et al. 2010). HGT experiments in animals have so far failed to produce results that would justify performing gene transfer in humans; however, gene transfer techniques may eventually be used for enhancement purposes. While any existing animal sensory modality suggests a possibility for future enhancement, visual enhancements are of particular interest (Jacobs et al. 2007). An inherent plasticity in the mammalian visual system may permit the emergence of a new dimension of sensory experience based solely on gene-driven changes in receptor organization. However, it should be noted that sensory systems are in large part integrated in the central nervous system. We should not expect to get a bird's vision by simply implanting genes that alter the form of the eyes. To benefit from bird's visual systems, we also need some of the neurological hardware that birds have and that we lack. This kind of gene transfer is therefore much more complicated and advanced than that which allows the zebra fish to glow. Keeping that in mind, I will now explore some possible sensory modalities that could be of interest for researchers to consider.

Although human vision is quite good when compared with that of other mammals, the visual systems that some birds enjoy are simply formidable. There are two sorts of light receptors in the bird's eye, rods and cones. Rods, which contain the

visual pigment rhodopsin, are better for night vision, because they are sensitive to small quantities of light. Cones detect specific colors (or wavelengths) of light, so they are more important to color-orientated animals such as birds. Most birds are tetrachromatic, i.e., in possession of four types of cone cells, each with a distinctive maximal absorption peak. In some birds, the maximal absorption peak of the cone cell responsible for the shortest wavelength extends to the ultraviolet (UV) range, making them UV-sensitive. Birds can also resolve rapid movements better than humans, for whom flickering at a rate greater than 50 Hz appears as continuous movement. This means that while humans cannot distinguish individual flashes of a fluorescent light bulb oscillating at 60 Hz, birds like budgerigars and chickens have flicker thresholds of more than 100 Hz. Birds can also detect slow moving objects. The movement of the sun and the constellations across the sky is imperceptible to humans, but detected by birds. The ability to detect these movements allows some migrating birds to properly orientate themselves (Jones et al. 2007).

Many animals have better night vision than humans, the result of one or more differences in the morphology and anatomy of their eyes. These include having a larger eyeball, a larger lens, a larger optical aperture (the pupils may expand to the physical limit of the eyelids), and more rods than cones (or rods exclusively) in the retina. Among other animals, cats and dogs have a layer of tissue in the eye that reflects visible light back through the retina. This tissue, referred to as “tapetum lucidum,” increases the light available to the photoreceptors, allowing the animal to see in poor light conditions.

Some animals have been known to perceive infrasonic waves going through the earth caused by natural disasters and can use these as an early warning. Infrasound is also used for long-distance communication by many of the large mammals. For example, elephants produce infrasound waves that travel through solid ground and are sensed by other herds using their feet, although they may be separated by hundreds of kilometers (Payne et al. 1986). These calls range from 15 to 35 Hz and can be as loud as 117 dB (Langbauer et al. 1991). Other animals are able to perceive ultrasound. For example, bats can detect frequencies beyond 100 kHz, possibly up to 200 kHz (Popper and Fay 1995).

Echolocation is a biological sonar used by bats, dolphins, and other animals. They use echolocation by emitting sounds out to the environment and listening to the echoes that return from the various objects. The animal can measure the range to the objects surrounding it by measuring the time delay between the animal’s call and the echo. The relative intensity of the sound and the difference in time delay between the animal’s ears gives the animal an idea of the horizontal angle of the object. Echolocation requires, in addition to the ability to emit sounds at high frequencies, also the ability to construct a detailed representation of a complex environment from sound. Thus, although echolocation does not require new sensory modalities, it requires neuronal structures alien to the human brain (Jones 2005).

Electroreception is the ability in some animals to sense electrical stimuli. It has been only been observed in aquatic or amphibious animals, since water is a much

better conductor than air. Electroreception is used in electrolocation (detecting objects) and for electrocommunication. Passive electroreception relies upon ampullary receptors which are sensitive to low (below 50 Hz) frequency stimuli (Collin and Whitehead 2004).

While these sensory modalities might prove very difficult to transfer to a human organism, the mere fact that they are possible in species that are relatively similar to us suggests that it is possible. We should also distinguish these *biologically* possible sensory modalities with other, merely *physical* possibilities. While it might be consistent with the laws of physics to be able to perceive neutrinos, this is not likely to happen for any organism that is even remotely human. In comparison, the biological changes for some of these sensory modalities are quite modest.

Should We Accept Sensory Enhancement?

Although human enhancement has been part of the bioethics debate for more than a decade, the issue of sensory enhancement has garnered little attention. A possible explanation may be that sensory enhancement seems to yield small benefits in contrast with, for example, cognitive enhancement. The claim that the benefits of sensory enhancement will at best be modest seems plausible. However, some justification for this kind of human enhancement is still possible to formulate. In this final section, this chapter will sketch out three arguments in favor of sensory enhancement, all of which need to be explored further.

Instrumental Value: To Perceive New Value Structures

The English expression "Art for art's sake" expresses the idea that art has intrinsic value, separate from whatever utility it may provide. This view is often defended by philosophers who ascribe to pluralist notions of intrinsic value, such as William Frankena and others (Frankena 1973). Some forms of artistic expression such as music, photography, and cuisine are directed to one or more specific sensory modalities such that these senses are necessary to experience the aesthetic beauty or excellence conveyed by these expressions. Sensory modalities are thus instrumental for us to be able to perceive or experience these sources of value. If some sensory enhancements became widespread, art that would require these enhancements to be perceived might be produced. Ultrasound musical instrument is one possible example. Although some sensory modalities are of little practical benefit, such as the ability to perceive UV light, these enhancements may be instrumental for us to experience and express new forms of human creativity. Although the deaf community can plausibly make the case that being unable to hear does not imply that one's life contains less welfare than otherwise, they cannot deny that being deaf deprives the person from experiencing a certain form of human expression that is, if not intrinsically valuable, at least very enjoyable (Cooper 2007). Music as such is

not the only possible form of expression, and we may all be “deaf” with respect to the as of yet unexplored and unimagined forms of cultural activity that may involve nonhuman sensory modalities. The argument against deafness can thus be deployed in favor of extending the range of human sensory modalities. It is worth noting that this argument is not specific to sensory enhancement. For example, some forms of cognitive enhancement may allow us to appreciate some literary works or the aesthetic qualities in abstract mathematics.

However, it is not obvious that the analogy between music and a hitherto unknown but possible source of intrinsic value can be made. Whether this analogy is adequate depends on the source of value of music and other existing art forms. Does the value of art rest on the communitarian importance of participating and sharing an experience across time and cultures? Or does the value of art depend on its intrinsic aesthetic qualities? An art form based on nonhuman sensory modalities would not in the same sense carry these communitarian values. However, some art theoreticians argue that art has cognitive value, in virtue of its capacity to increase our understanding on some topic (Schellekens 2007). Furthermore, these hypothetical nonhuman art forms may be just as beautiful and aesthetically pleasing as existing art forms. A similar argument in favor of expanding sensory modality has been formulated by Nick Bostrom. He argues that although our human limitations are so pervasive that we fail to notice them, there is likely to be a huge range of modes of being in and perceiving the world that allow us to engage in very valuable activities. It is therefore important to explore these possible modes of being (Bostrom 2003).

Two Arguments from Diversity

The human species is extraordinarily homogenous in comparison with other primate species, which often include a number of subspecies. It has been hypothesized that this relative homogeneity may be due to a great disaster in our recent evolutionary history (Ambrose 1998). This homogeneity may prove problematic. Diverse societies are more innovative and more adept to understand disruptive social or natural dynamics (Kandler and Laland 2009). Homogenization may expose us to risks if our limited frames of reference reduce our ability to understand and imagine possibilities and risks. Remember that for many unexpected catastrophic events, it is our failure to anticipate these events that make them so dangerous (Taleb 2007). By increasing diversity among the constituent members of our global civilization, we may increase the possibility to imagine and anticipate some of the risks that we have as of yet failed to conceive of. Sadly, the modern condition seems to reduce the diversity of opinions, values, and beliefs. Urbanization, globalization, and a truly global commercial culture have already led to the elimination of many languages and cultures (Abrams and Strogatz 2003). As mankind will become predominantly a species of city dwellers in the twenty-first century, this process will probably accelerate. As sensory modalities constitute

literally a new perspective on the world, the proliferation of different combinations of perceptual apparatuses may contribute to a significant degree to different ways of relating to and interfacing with the world. As has been discussed in this chapter, many of the sensory modalities that may become possible include trade-offs. There is no "perfect sense," or perfect combination of enhancements that would be great for anyone in any context. For example, night vision may make a person react aversely to strong light or rapid changes in light conditions. Accurate hearing may sound great, but may cause stress and anxiety in loud environments. It is therefore unlikely that everyone will opt to implement the same enhancements.

Is there an intrinsic value in diversity? Some philosophers and bioethicists tend to believe so, and this author is inclined to share that appreciation. Certainly, many opponents of human enhancement fear that human enhancement may bring about the homogenization of the human species (Kass 2002). These authors seem to ascribe intrinsic value to the prevailing level of diversity. However, it is worth reflecting on whether or not mankind happens to find itself at an optimal (from both an intrinsic and instrumental perspective) level of diversity. As this seems unlikely, the intuition that we happen to be at a local optimum with regard to diversity may reflect a status-quo bias, as argued by Nick Bostrom and Toby Ord (Bostrom and Ord 2006). Perhaps *increased* diversity is morally desirable. Sensory enhancement could bring about this diversity.

The Intrinsic Value of Sense Modalities

Amartya Sen and Martha Nussbaum formulated in the 1980s a view of human welfare that has become influential in welfare economics, referred to as the capabilities approach (Sen 1989). According to this view, human welfare is not possible to reduce to subjective well-being, but neither can it be plausibly measured in resources (Sen 1993). Rather, it is what we can accomplish with these resources that matters. According to the capabilities approach, there are some functional capabilities that are good for the person who has them, and people deprived from these capabilities are worse off even when the deprived person does not want or cannot imagine having these capabilities. Imagine, for example, a slave working in a mud brick factory. This person may believe that it would be bad for him or her to be able to read. This does not make the slave less deprived, according to the defendants of this view (Nussbaum 2000). Functioning includes bodily health, bodily integrity, and being able to use one's senses and so on (Sen 1993). Could this framework be extended to include human enhancement? Although this is not what the proponents of the capabilities approach intended, the definition of "normal" health, perceptual ability, and limb functionality seems to be highly context-sensitive. In modern welfare-states, it is a sign of deprivation to lack teeth in one's late fifties, something that was normal 100 years ago. Are we in a similar sense deprived because we lack the sensory modalities that might be available to future generations? If the capabilities approach is correct, the mere fact that we do not regret lacking night vision does not imply that we are not deprived.

Conclusion

Sensory enhancement has long been neglected in the debate on human enhancement. This chapter has argued that sensory enhancement matters and that there are reasons to allow people to experiment with it. Since the technological feasibility of some sensory enhancements is quite tangible, and that the first sensory enhancement devices may reach the mass market in only a few years, there is an evident need for a discussion on the normative issues involved.

Cross-References

▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethical Implications of Cell and Gene Therapy
▶ Ethical Implications of Sensory Prostheses
▶ Ethical Issues in Cochlear Implantation
▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
▶ Gene Therapy and the Brain
▶ Reflections on Neuroenhancement
▶ Research in Neuroenhancement

References

Abrams, D. M., & Strogatz, S. H. (2003). Modeling the dynamics of language death. *Nature, 424*, 900.

Ambrose, S. H. (1998). Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *Journal of Human Evolution, 34*(6), 623–651.

Ashton, K. (2009). That 'Internet of things' thing. *RFID Journal*, http://www.rfidjournal.com/article/view/4986. Accessed 8 April 2011.

Bostrom, N. (2003). Transhumanist values. In F. Adams (Ed.), *Ethical issues for the 21st century*. Oxford: Philosophical Documentation Center Press.

Bostrom, N., & Ord, T. (2006). The reversal test: Eliminating status quo bias in applied ethics. *Ethics, 116*, 656–679.

Collier, R. (2010). Rosy outlook for people with diabetes. *Canadian Medical Association Journal, 182*(5), E235–E236.

Collin, S. P., & Whitehead, D. (2004). The functional roles of passive electroreception in non-electric fishes. *Animal Biology, 54*(1), 1–25.

Cooper, R. (2007). Can it be a good thing to be deaf? *Journal of Medicine and Philosophy, 32*(6), 563–583.

Frankena, W. (1973). *Ethics* (2nd ed., pp. 87–88). Englewood Cliffs: Pearson.

Goldman, J. (2012). Google Glass Explorer Edition. *Cnet*. http://reviews.cnet.com/camcorders/google-glass-explorer-edition/4505-9340_7-35339166.html?ttag=fbwp

Hameed, J., Harrison, I., Gasson, M. N., & Warwick, K. (2010). *A novel human-machine interface using subdermal magnetic implants*. Proceedings IEEE International Conference on Cybernetic Intelligent Systems, Reading, pp. 106–110.

Hickey, H. (2008). Bionic eyes: Contact lenses with circuits, lights a possible platform for superhuman vision. *The UW Faculty and Staff Newspaper, 25*(12). University of Washington.

Jacobs, G. H., Williams, G. A., Cahill, H., & Nathans, J. (2007). Emergence of novel color vision in mice engineered to express a human cone photopigment. *Science, 315*(5819), 1723–1725.

Jones, G. (2005). Echolocation. *Current Biology, 15*(13), R484–R488.

Jones, M. P., Pierce, K. E., Jr., & Ward, D. (2007). Avian vision: A review of form and function with special consideration to birds of prey. *Journal of Exotic Pet Medicine, 16*(2), 69–87.

Kandler, A., & Laland, K. N. (2009). An investigation of the relationship between innovation and cultural diversity. *Theoretical Population Biology, 76*(1), 59–67.

Kass, L. (2002). *Life, liberty, and defense of dignity: The challenge for bioethics* (p. 48). San Francisco: Encounter Books.

Langbauer, W. R., Payne, K. B., Charif, R. A., Rapaport, L., & Osborn, F. (1991). African elephants respond to distant playbacks of low-frequency conspecific calls. *The Journal of Experimental Biology, 157*(1), 35–46.

Lingley, A. R. et al. (2011). A single-pixel wireless contact lens display. *Journal of Micromechanics and Microengineering, 21*(12), 125014 (8pp).

Nagel, S. K., Carl, C., Kringe, T., Märtin, R., & Konig, P. (2005). Beyond sensory substitution – Learning the sixth sense. *Journal of Neural Engineering, 2*, R13–R26.

Norton, Q. (2006). A sixth sense for a wired world. *Wired Magazine*. http://www.wired.com/gadgets/mods/news/2006/06/71087

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach*. Cambridge: Cambridge University Press.

Payne, K. B., Langbauer, W. R., & Thomas, E. M. (1986). Infrasonic calls of the Asian elephant (*Elephas maximus*). *Behavioral Ecology and Sociobiology, 18*(4), 297–301.

Popper, A., & Fay, R. R. (Eds.). (1995). *Hearing by bats* (Springer handbook of auditory research, Vol. 5). New York: Springer.

Romer, A. S., & Parsons, T. S. (1986). *The vertebrate body* (6th ed.). Philadelphia: Saunders College Publishing.

Schellekens, E. (2007). *The aesthetic value of ideas. Philosophy and conceptual art*. Oxford: Oxford University Press.

Sen, A. (1989). Development as capability expansion. *Journal of Development Planning, 19*, 41–58.

Sen, A. (1993). Capability and well-being. In M. Nussbaum & A. Sen (Eds.), *The quality of life* (pp. 30–53). New York: Oxford Clarendon Press.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Walcott, C. (1996). Pigeon homing: Observations, experiments and confusions. *The Journal of Experimental Biology, 199*(Pt 1), 21–27.

Wilson, A. D., & Baietto, M. (2009). Applications and advances in electronic-nose technologies. *Sensors, 9*(7), 5099–5148.

Zhiyuan, G., Jiangyan, H., Bensheng, J., Toong, J. L., Yanfei, X., & Tie, Y. (2010). Chimeric gene constructs for generation of fluorescent transgenic ornamental fish. U.S. *Patent No. 7, 834, 239*. Alexandria, VA: National University of Singapore.

Section XI

Ethical Implications of Cell and Gene Therapy

Ethical Implications of Cell and Gene Therapy

52

Heiner Fangerau

Contents

Abstract

Stem cell implantation, brain tissue transplantation, and gene therapy for the brain share a common feature. On the one hand, they are seen as potentially beneficial in the quest to cure neurological disorders; on the other hand, ethicists and the public have concern about these interventions affecting a person's identity. This section introduction addresses the rationale behind these concerns and further ethical issues related to intracerebral cellular or genetic interventions. It provides a springboard for the following chapters by Christian Lenk, Heiner Fangerau and Norbert Paul.

Introduction: Identity

Stem cell implantation, brain tissue transplantation, and gene therapy for the brain share a common feature. On the one hand, they are seen as potentially beneficial in the quest to cure neurological disorders; on the other hand, ethicists and the public have concern about these interventions affecting a person's identity. This would be true if the brain is viewed as an organic correlate or the seat of one's identity.

H. Fangerau
Department of History, Philosophy and Ethics of Medicine, University of Ulm, Ulm, Germany
e-mail: heiner.fangerau@uni-ulm.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_52,

Basic theories of identity in this context encompass the biological, psychological, and narrative aspects of identity. The theories focus on the biological organism, psychological continuity, and the connectedness and incorporation of experiences into personal history (for an overview, see Shoemaker 2012. For implications of identity concepts on associated fields of interest like the evaluation of human-animal chimeras, see Badura-Lotter and Fangerau 2014). There is concern that stem cells, genetically altered cells, or gene carriers inserted into the brain could interact with native brain cells in a way that fundamentally alters the person's identity. Furthermore, these changes could be irreversible due to the nature of the transplanted cells or genetic modifications. Changes in identity might ultimately be seen as harmful because they essentially represent the end of the original person (Goldstein 2013). This raises significant questions of whether it is morally acceptable to attempt these types of therapies.

The assumed alterations to a person's identity must be examined to determine if these therapies comply with ethical principles in medicine like beneficence and non-maleficence. The target diseases for these therapies are mostly degenerative neurological diseases or diseases resulting in loss of brain tissue such as Parkinson's disease, stroke, different types of dementia, or even psychiatric disorders. Because these diseases can produce personality changes themselves, personality change following treatment is not generally thought to be harmful (Goldstein 2013). To what extent personality changes are acceptable to patients and their families and how severe they would need to be before they would be considered fundamental changes to the patient's identity need to be investigated.

Intention, Control and Ontological Questions

Additionally, the intentions behind stem cell or gene therapies should be evaluated from a moral perspective. The public generally accepts therapeutic testing with the intention to treat diseases. However, testing to simply satisfy curiosity, enhance reputation, or make money is viewed negatively. Therefore, scientific publications on transplanting brain tissue or stem cells have emphasized the therapeutic intentions of authors. The respective scientists aim to develop regenerative therapies for replacement or regeneration of neural tissue. Their reasoning is based on a mechanistic conception of medicine in general and brain physiology in particular (Fangerau 2011). Furthermore, their reasoning is based on the idea that they can control the consequences of their therapeutic actions. The need to control therapies is linked to this understanding of causal therapies. If transplanted stem cells could be restrained and controlled to only act as "pumps" that secrete growth factors to enhance regeneration of endogenous cells or supplement neurotransmitters, they may be viewed differently than transplanted cells that can connect or interact with, invade, or possibly overtake the original brain cells (Fangerau and Trapp 2011). Controllable transplants can be removed if necessary and possible harmful effects can be limited. Uncontrollable cell transplants run risks of growing in unrestrained

ways that would lead to unforeseen, lasting, and progressive complications. The inability to control transplants has proven to be a real danger in the past. As a consequence intention and controllability have become interconnected touchstones in the ethical debates surrounding neuro-regeneration and neurogenetics.

Additional ethical issues exist with stem cell transplantation and gene therapy in the brain. Besides the general consideration about interfering with the personality of the patient, the ontological status of the transplanted tissue is of moral relevance. Some transplants originate from human embryonic tissue. This has generated significant debate over the moral legitimacy of utilizing potential human life to developing cures and the possibility of promoting abortion (see ▶ Chap. 53, "Neural Transplantation and Medical Ethics: A Contemporary History of Moral Concerns Regarding Cerebral Stem Cell and Gene Therapy" by Fangerau and Paul). The status of the transplanted or genetically altered cell should also be considered. Appropriate ethical evaluation requires classifying transplanted tissues as parasitic (they take over the host brain), symbiotic (they cooperate with the host brain), or autonomous (they behave both ways) (Bührle 2011). If potential therapies are to be tested in animals before they are used in human clinical trials (a contested issue itself; see Mauron and Hurst 2011), then implanting human brain cells and of human genes in animals creates chimeras and challenges ethical values like species integrity and human dignity (Clausen 2011). Also, testing novel neuronal stem cell and gene therapies in animals has been debated since the inception of the idea. Finally, fears about the lack of control over therapies after administration have raised questions about research ethics and the potential abuse of study subjects, especially, because potential subjects for human trials might be unable to provide full informed consent to studies if they suffer from degenerative neurological diseases and have reduced cognitive function. General issues like sham surgeries or the use of placebos should be considered with novel intracerebral interventions as well.

Conclusion and Future Directions

The following chapters by Christian Lenk, Norbert Paul, and Heiner Fangerau address these issues in depth and discuss the relatively short history of these two types of intracerebral therapies along with the basic moral values in medicine that they challenge. Ultimately, evaluating the risks and benefits of these therapies will have to consider previous evidence and fundamental aspects of the brain, mind, and identity to understand the ethical challenges that each presents.

Cross-References

▶ Ethics of Sham Surgery in Clinical Trials for Neurologic Disease
▶ Gene Therapy and the Brain

▶ Impact of Brain Interventions on Personal Identity
▶ Informed Consent and the History of Modern Neurosurgery
▶ Neural Transplantation and Medical Ethics: A Contemporary History of Moral Concerns Regarding Cerebral Stem Cell and Gene Therapy
▶ Neuroethics and Identity
▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives

References

Badura-Lotter, G. & Fangerau, H. (2014). Human–Animal Chimeras: Not Only Cell Origin Matters. *American Journal of Bioethics, 14*(2) (pp. 21–22).

Bührle, C. P. (2011). Changes in personality: Possible hazards arising from stem cell grafts – An ethical and philosophical approach. In H. Fangerau, J. M. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 57–89). Bielefeld: Transcript.

Clausen, J. (2011). Establishing regenerative medicine for the human brain: Ethical aspects of intracerebral stem cell transplantation. In H. Fangerau, J. M. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 91–106). Bielefeld: Transcript.

Fangerau, H. (2011). Brain, mind and regenerative medicine: Ethical uncertainties and the paradox of their technical fix. In H. Fangerau, J. M. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 15–30). Bielefeld: Transcript.

Fangerau, H., & Trapp, T. (2011). Introduction. In H. Fangerau, J. M. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 7–12). Bielefeld: Transcript.

Goldstein, J. (2013). *Personale Identität und intrazerebrale Stammzelltransplantationen*. Dissertation Medizinische Fakultät Universität Ulm.

Mauron, A., & Hurst, S. (2011). Experimenting innovative cell therapies for Parkinson's disease: A view from ethics. In H. Fangerau, J. M. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 107–122). Bielefeld: Transcript.

Shoemaker, D. (2012). Personal identity and ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2012 Edition). http://plato.stanford.edu/archives/spr2012/entries/identity-ethics/. Retrieved October 27, 2013.

Neural Transplantation and Medical Ethics: A Contemporary History of Moral Concerns Regarding Cerebral Stem Cell and Gene Therapy

53

Heiner Fangerau and Norbert W. Paul

Contents

Abstract

The possibilities of brain transplantation, neural transplants, and neural grafting as well as early approaches of gene therapies have inspired the fantasies and optimism of both scientists and the public. They have also caused hopes and fears at the same time, which are reflected in ethical debates about their use and usefulness in medicine. Based on the scientific and public discourse about neural transplants and genetic interventions in a broad understanding, the aim of this chapter is to reconstruct the major aspects of the development of a medical technology and the ethical concerns, which have accompanied its application since the 1890s. Ethical debates may vary in different countries based on different cultures and traditions. Therefore, we must stress that the main focus of this chapter is based on the English and German literature.

H. Fangerau (✉)
Department of History, Philosophy and Ethics of Medicine, University of Ulm, Ulm, Germany
e-mail: heiner.fangerau@uni-ulm.de

N.W. Paul
History, Philosophy, and Ethics of Medicine, Johannes Gutenberg University Medical Center, Mainz, Germany
e-mail: npaul@uni-mainz.de; manuela.sackissow@uni-mainz.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_127,

Introduction: The Dystopia of Brain Transplantation

The term "neural transplantation" covers a wide range of possible applications within the brain. Complete brain transplantations, homoplastic replantation, transplantation of parts of the brain, brain cell transfer, transfer of genetically manipulated cells, the implantation of fetal neural tissue, and intracerebral stem cell application are covered by this term. Over the course of history, a great variety of methods for interfering with the brain at a cell based level have been used. A comprehensive review by Gopal Das described the replantation of adult neural tissue, implantation of fragments of peripheral nerves, implantation of spinal ganglia, transplantation of nonneural tissue, implantation of tumors, and transplantation of embryonic neural tissue as methods that have been explored between 1890 and 1990 (Das 1990). However, until the late 1980s, the public only seemed to have substantial reactions to reports of complete brain transplantations or brain grafting, which could be interpreted as a major step toward routine brain transplantations, or to put it differently, body transplantations, if the brain is conceived as the host of the "self."

At the time, skepticism was rampant and the horrifying aspects of changing a person's self by transplanting the brain was a timely topic of discussion. At its best, this combination of concerns was reflected in a newspaper caricature, which was possibly triggered by Gilman Thompson's report on "successful brain grafting" in 1890 after he had transplanted pieces of cortical tissue from adult dogs and cats into the brains of recipient dogs in order to test the vitality of transplanted neural tissue in a series of experiments (Thompson 1890). Thompson's goal was not to reconstruct damaged brains, but rather to test the degeneration of brain cells under foreign conditions following evolutionary concepts of the atrophy of non-used organs. While Thompson himself and the reception of his experiments in the journal *Science* was cautious with regard to the practical value (Anonymous 1890), other reactions to his findings reinterpreted or exaggerated the goals and outcomes of his experiments. In an article in the New York Times from 1891 under the heading "Brain Transplantation," the author recounts a conversation with an "amateur professor" who describes how he had performed an experiment in Germany as a "student under Haeckel" during which he had interchanged the cerebellums of a "Dachshund of the sort that Prince Bismarck is so fond of and an enormous tom cat" (J. D. 1891). In a satirical manner, the professor describes his experience in semi-medical terms and in the style of a journal article of that time using the tunneled perspective of an experimenter ("... made an incision in the occiput of each, after both had been carefully washed in a weak solution of bichloride of mercury to prevent any harm being done by noxious bacilli, bacteria, and such things..."). At the end of the story, which is full of references to German names, terms, and stereotypes, the dog behaves like a cat and the cat behaves like a dog. After both animals were sacrificed, the experiment was successfully repeated by exchanging the cerebrums. Without mentioning Thompson and his experiments directly, this article reflects the public reception of his studies: On one hand, the author appears very skeptical with regard to their outcome, as the fictional story

clearly states the perceived implausibility of successful brain grafting; on the other hand the author raises fears providing fictional consequences for the “self” of the cat and the dog, as both changed their behavior according to the brain and not their body. The story is settled in a German experimental context of the late 1890s with clear references to horror stories about the new science of experimental physiology in the style of Marry Shelley’s Frankenstein.

This idea of neighboring brain transplantations with Dr. Frankenstein’s artificial creation of life survived throughout the twentieth century. During the late 1960s, the neurosurgeon Robert J. White reported to have transplanted monkey heads (and brains) to monkey bodies, and in the late 1970s, he postulated on performing these procedures in humans. Because of these experiments, which were broadcasted on German television, the media called him “Frankenstein of Ohio” (Anonymous 1976), a recurrent motif in any discussion about brain transplantation since then (and before). White himself also referred to Frankenstein in an article from 1992 when he criticized the transplantation of human fetal brain tissue into the brains of patients suffering from Parkinson’s disease in order to cure their disorder. He was not only skeptical about the possible clinical results, but he also disapproved of the killing of fetuses for harvesting fetal brain tissue (White 1992). By that time, however, White’s experiments did not form part of the scientific discourse any longer. They were only remembered in news magazines or online news as horrifying stories of the past (Widmann 2000) and “scientifically misleading” (Anonymous 2001).

The Use of Fetal Tissue

While complete brain transplantations remained science fiction through the twentieth century, as the heart surgeon Christian Barnard claimed according to newspaper reports in 1968 (“I do not think a head or brain transplant is possible in our lifetime or even the lifetime of our children”) (Anonymous 1968), the use of fetal brain tissue for brain transplants in order to cure Parkinson’s disease started to become a reality in the 1980s. This “‘modern era’ of brain tissue transplantation,” as the neuropsychologist William Freed put it (Freed 2000, p. 32), not only changed the prospect of neural transplantation toward the hope for regular clinical application, but also brought about a new ethical debate of a different quality that added to the question of the “self” in brain transplantation questions surrounding the use of fetuses and clinical trials with neural tissue.

After an initial peak of interest in various neural tissue transplantation types in the late nineteenth and early twentieth centuries, only sporadic papers had been published on this issue. Between 1946 and the late 1970s, the Science Citation Index lists less than a handful of papers per year on the topic. However, in the late 1970s and especially after 1982, this field of medical research experienced a renaissance, which Gopal Das called a “mushrooming of research activity” (Das 1990). The frequency of reports on successful grafts of various kinds accelerated. The early 1980s can be considered as a watershed, when Low et al.

published a paper in *Nature* in 1982 on "Functional recovery following neural transplantation of embryonic septal nuclei in adult rats with septohippocampal lesions" (Low et al. 1982), which in the media was seen as a promising starting point for "Transplant hopes for brain tissue" (Anonymous 1982). That same year, Backlund et al. were experimenting with the transplantation of material from the adrenal medulla of patients suffering from Morbus Parkinson into their caudate nucleus, hoping that the dopamine produced by this tissue might improve the symptoms of the disorder (Backlund et al. 1985). While studies in rodents using fetal tissue initiated a research surge that promoted the use of fetal tissue as an almost ideal material for therapeutically intended brain transplants, the studies on autografts with adrenal tissue opened the frontier of clinical trials using brain transplants in patients. Fetal tissue grafting was seen as a promising approach, because fetal tissue lacked "extensive neural processes," was "less susceptible to anoxia than adult animal neurons," maintained viability after cryopreservation, making it easier to store, lacked "major histocompatibility complex (MHC) surface markers, offering relative protection to immunologic rejection," and was "capable of producing a number of trophic factors, which may preclude the need for cografting" (Boyer and Bakay 1995, p. 116).

From Bench to Bedside?

While some public media such as the *Economist* remained cautious at first regarding the therapeutic prospects of these studies, it did not take long until the first clinical trials were performed during which patients received fetal allografts. Quoting the above mentioned rodent experimentalists, the *Economist* stated "that much remains to be done before human brain transplants can become accepted therapy" (Anonymous 1983, p. 86). However, after a limited number of studies with medullary autografts during the late 1980s, the early 1990s saw a reasonable number of reports on neural transplants in humans using fetal material (Boyer and Bakay 1995), and scientists were pushing neural transplantation as an established field of research that needed to be translated into practical therapeutic medicine as a remedy for various degenerative disorders.

Overviews of the history of neural transplantation have been published by well-informed experimentalists and specialists of the field (Gash 1984; Björklund and Stenevi 1985a; Gash et al. 1985; Borges 1988; Das 1990). With the help of these articles, the authors created a narrative of a research tradition that has more or less moved in a straight direction forward from Thompson's first reports mentioned above until the research of that time (occasionally literally in a timetable), sometimes explicitly excluding studies by Robert White (Freed 2000, p. 38). These authors placed themselves and their research in a serious tradition of experimental work, which they used to challenge reports by surgeons like White as not serious and sensation-seeking. During this time, several monographs on neural grafting and transplantation were published, including conference proceedings and other edited volumes (Wallace and Das 1983; Björklund and Stenevi 1985b; Das 1986; Gash and

Sladek 1988; Lindvall et al. 1991; Dunnett 1992). Following the concept of Ludwik Fleck regarding the manifestation of knowledge, during these years, the practice of brain transplantation was shifted by protagonists from the level of a new research frontier in journal literature to the more general knowledge of textbooks (for an overview see Martin and Fangerau 2006). A second wave of published textbooks followed at the end of the 1990s, which represented the "decade of the brain" as termed by the US Congress for the period from 1990 to 1999 as an endeavor "to enhance public awareness of the benefits to be derived from brain research" (Jones and Mendell 1999).

Ironically, neurosurgeons had started to hint at the clinical potential of this new research shortly after 1982. From the perspective of sociology, their appeal to the neurosurgical community to participate in this new field of research and application might have been triggered by the perception that the first promising experiments had been performed by a team of physiologists and psychologists. Now, it was felt that the clinical application should be handed over to neurosurgeons. As early as 1983, the official journal of the Congress of Neurological Surgeons, *Neurosurgery*, published a paper by neurosurgeon James E. Wilberger on the "Transplantation of Central Nervous Tissue" in which he repeatedly stated that "although the majority of research in CNS transplantation" was in the hands of neurobiologists, "the information generated thus far is closely approaching the point of making CNS transplantation a clinical tool for the hands of the neurosurgeon" (Wilberger 1983, p. 93). Similarly, seven years before his critical remarks about the use of fetal tissue, the notorious Robert White had expressed his hope in an editorial to the journal *Surgical Neurology* that the neurosurgical community would actively participate in this expanding field and be involved in a "new dimension of neurosurgery," "if and when this intriguing and fascinating line of neurological investigation becomes of human significance" (White 1985).

Early Genetic Interventions into the Brain as a Stairway to a New Paradigm of Control

It was also in the late 1970s and 1980s, when the rise of novel technologies for the manipulation of genes, especially the introduction of restriction enzymes and recombinant DNA, paved the road toward genetic interventions into the human brain. Historically and systematically, genetic interventions were developed along the routes of cancer research and molecular biology. At the same time, environmental, viral, and genetic paradigms used to explain cancer were amalgamated in the "unifying pathway of cancerous growth" (Fujimura 1992) after Bishop and Varmus developed and popularized the so-called oncogene paradigm (Bishop 1982), translating former models of carcinogenesis into cancer genetics. Gene therapy came into sight as a tool to correct cancerous growth by inserting corrective genetic material into cells and tissues which lost their ability to counterbalance the proto-oncogenic potential which was now understood to be part of the genetic inventory of every "normal" cell. Four major obstacles had to be overcome in order

to come to grips with the concept of gene therapy: (1) identification of the dysfunctional gene(s); (2) targeting of those genes with corrective material; (3) integration of the corrective genetic material in the affected cells or tissues; (4) controlled expression of the corrective genes in order to contribute to "normal" cellular behavior. Against this background, it is understandable, why first protocols for gene therapy were dealing with diseases of the blood. The targeting of cells in the blood by simply infusing corrective materials – with or without lipid-packaging or viral vectors – was certainly more easily accomplished than the targeting of cells in tissues like those of the central nervous system.

First attempts of gene therapy resulted in public concern, such as the controversial experiments of Martin Cline who treated patients with beta-thalassemia in non-authorized clinical trial in 1980 and was persuaded to resign from his position at UCLA after the experiment went public. This first episode of gene therapy highlights the ethical and regulatory uncertainties at the biomedical frontier and as a reaction, first regulations with regard to clinical trials involving recombinant genetic materials were put into place in the United States in 1985, when the National Institutes of Health issued guidelines particularly addressing the issue of genetic alteration. Gene therapy, which was regarded to be too premature for clinical trials, was brought back from bed to bench-side. This resulted in a situation in which the flourishing basic research was more and more regulated, but neither regulators nor the public were prepared for the recurrence of gene therapy. This first successful gene therapy was performed in a young girl suffering from a congenital severe combined immune deficiency (ADA-SCID). The protocol was presented and debated in numerous public hearings and ethics committees and the editor in chief of the just founded journal "Human Gene Therapy" French Anderson resumed "And yet, in spite of all this media attention, a considerable portion of the public, including large segments of the medical and scientific communities, did not really believe that an approved clinical attempt at gene therapy was going to take place for, perhaps, years. After the protocol was initiated and the gene-corrected cells were infused into the patient, the question we were asked over and over was: How did this happen so unexpectedly? The answer is, that society had simply not come to terms with the concept that genetic engineering was ready to begin" (Anderson 1990, p. 371f.)

As a consequence, the US authorities decided to establish a more standardized procedure for the evaluation of basic and clinical research dealing with recombinant genetic material. However, the system which included the reactivation of the recombinant advisory committee (RAC) established by the NIH in 1974 as a reaction to public concerns regarding the proliferation of recombinant materials into the natural environment failed its test when the first clinical application of gene therapy into the human brain was performed:

In 1992, the 52-year-old Clemma Hewitt in San Diego, California, was diagnosed with a glioblastoma, a malignancy of the human brain. After surgical intervention, radiation and chemotherapy, and an experimental approach using radioactive antibodies, the patient had to undergo surgical revision because of a recrudescence of the tumor. At this time, the University of California San Diego was developing a therapy based on altered DNA of the tumor itself, that is, the direct application of

recombinant DNA in proliferative tissue to induce "cellular suicide." This experimental approach was far from a clinical protocol, which would have passed the authorities. The physician of the patient, influential Ivor Royston, addressed the NIH using political pressure so that Officials at the National Institutes of Health waived the usual reviews for the new treatment for her. Researchers genetically altered Mrs. Hewitt's own tumor cells to give them a gene for interleukin-2, a substance involved in the production of white blood cells. Her doctors hoped that the altered cells would produce a type of white blood cell that would attack the cancer.

This brief narrative illustrates at least two aspects: First, genetic interventions into the brain were regarded as a last resort applicable in a compassionate release of an experimental protocol with reasonable shortcomings regarding the proof of safety, specificity, and efficacy. Second, the fact that the procedure of compassionate release of premature clinical protocols was established as a standard exception at the NIH in 1993 based on the Clemma Hewitt case might be indicative for the fact that again, neither the scientific community, the medical field, the authorities, nor the public had "come to terms with the concept" that the application of recombinant genetic material into the human brain was ready to begin (Paul 2003).

Long since, genetic modifications not only target malignant growth in the human brain but venture into spheres where cognition and its biological basis in the human brain are inextricably intertwined. A number of studies addressed specific clinical phenotypes and the potential not only to better understand diseases like Alzheimer's disease based on genetics and genomics, but also target or prevent the disease by genetic alteration (Carlson et al. 1997).

Ethical Considerations Regarding Human Brain Grafting

Similarly, skepticism and ethical concerns regarding human brain grafting with fetal tissue were significantly articulated, especially after the human significance of these approaches had vehemently come into sight or had already become reality. For example, Karen Morrison raised some severe ethical considerations in a careful review of the state of the art of brain transplantation, in which she referred to Backlund's experiments, explained the techniques of implanting cells, described the limited primate experiments, and discussed both the realm of connected therapeutic strategies as well as the point of intervention (at an earlier or later stage of Parkinson's disease, for example). Under the heading "Brain transplantation – still fantasy?" she asked (in the order of appearance) whether there was a "moral distinction ... between fetal pancreatic transplantation and fetal brain transplantation," whether there was "enough yet known about transplantation in primates to proceed to human trials," whether therapeutic trials were ethical, whether "moral issues implicit in using fetal tissue" could be avoided using donor tissues "obtained by patient consent" and finally, which "profound ethical questions" were raised by using cross-species grafting (Morrison 1987, p. 443). Without delving deeper into these moral issues, Morrison fully captured most of the ethical issues that would be subsequently discussed in the following years with regard to brain grafting with fetal tissue as well as stem cell implantation after the year 2000.

On a more abstract level, these issues also included the question of possible alteration of personal identity by transplants, the formation of chimera, and the instrumentalization of human embryos. In addition, the justification of clinical trials was questioned due to the lack of sufficient animal experimentation and sufficient evidence of efficacy as well as limited controllability and minimal reversibility. The question of altering the human mind and personal identity by transplantation and of the ethics of species hybridization mainly remained theoretical during the early 1990s and again in connection with stem cell research after 2000 (Greene et al. 2005; Berg 2006; Wade 2005). They were mostly treated in the form of in thought experiments. However, the problems of embryo instrumentalization and experiments in humans gained practical importance very early on in the debate. In particular, the practical use of embryonic tissue raised concerns (Jones 1991). While scientists and some bioethicists had been quick to conclude "that retrieval of such tissue from fetal remains is analogous to the transplantation of organs or tissue obtained from adult human cadavers" (Mahowald et al. 1987), the question of the origin of this tissue raised concerns about induced abortions.

In the United States, these morally driven reactions had real consequences for scientists in the form of ethically based research restrictions at the end of the 1980s. When in 1987 a researcher proposed a research protocol to the National Institute of Neurological Disorders and Stroke (NINDS) in which he suggested to implant fetal neural tissue harvested after induced abortions into the brains of patients suffering from Parkinson's disease, the director of the NINDS passed it on to the director of the National Institutes of Health (NIH) to have it approved. The Institutional Review Board had not raised objections to the proposal, but the director of the NINDS had considered the protocol ethically critical enough to have it consented by the director. The director in turn sought advice from the United States Department of Health and Human Services (DHHS) (a cabinet department of the national government). As a consequence, the Assistant Secretary for Health issued a moratorium that stopped financial support from public federal sources for transplantation research using fetal tissue from induced abortions and established an advisory committee that was charged with discussing the controversial ethical issues (U.S. Congress 1990, p. 171; Hoffer and Olson 1991). Although the advisory board agreed that research with fetal tissue should be allowed under certain conditions, the ban was not lifted until 1993 by the then new Clinton administration (Clinton 1994). The moratorium had been under constant attack, because it was said to hinder research and possible cures for patients in the United States, while in "the meantime other countries, Sweden and Canada in particular, are making headway" (Beardsley 1990). Critics of the research ban conducted questionnaire surveys that were intended to research public attitudes regarding the ban, and concluded that "at least among psychology students, there is a favourable attitude to research on and clinical use of fetal neural tissue" (Sanberg 1990). At the same time, other authors articulated severe moral concerns regarding the use of induced abortions for gaining fetal transplants (McCullagh 1988). They saw abortion either as immoral *per se* or warned against the instrumentalization of possible life, fearing that women might become pregnant and donate their aborted fetuses without clear consent, or that (in line with the quoted

questionnaire survey) abortion would become socially acceptable if it was concomitantly connected to curing neurological diseases (Jones 1991; Coutts 1993). Thus, as a compromise, it was suggested that restrictions were needed "to prevent the use of grafts from encouraging induced abortions and to maintain high standards of respect for life and human dignity," but if these restrictions were enabled, then the use of fetal tissue to possibly cure disease could not be objected to anymore (Boer 1994, 1999). The hope and promise to find cures has been a strong argument for justifying morally ambiguous regeneration research before, and therefore, it is not surprising that this argument was also discussed and challenged in the discourse about neural transplantation. In addition, regeneration scientists tried to suggest technological solutions for moral objections (Fangerau 2011).

In a special issue of *Trends in Neurosciences* in 1991, Olle Lindvall explained the prospects of neural transplants in human neurodegenerative diseases such as Parkinson's, Huntington's, dementia, amyotrophic lateral sclerosis, and hereditary ataxia. He had been one of the first scientists who grafted tissue from the ventral mesencephalon of human fetuses to the brains of patients suffering from Parkinson's disease using a stereotactical method, but nevertheless, he carefully avoided calling neural transplantation a panacea in this article. While answering criticisms regarding the state and security of the technology as well as the associated fears to "try" an immature therapy that was possibly connected to neuroblastoma and behavioral changes in patients (Moss and Rosene 1985; Rogers et al. 1985), he maintained that neural grafting was still in its infancy and that clinical trials in a few, optimally selected patients would be necessary in the future. Like other careful authors of the period, he seems to have followed an ethical discourse brought up repeatedly in the discussion about neural transplantation on the ethics of communicating research results, which recommended not raising unrealistic hopes in patients by prematurely reporting results (Morrison 1987; McCullagh 1988; Rehncrona 1997). Thus, for Lindvall, only the "lack of adequate medical therapy" and the "severity of symptoms in these disorders" legitimized further research and clinical studies in patients, although the procedures in question had not been optimized before in animal experiments, just as it had been the case with premature protocols of gene therapy. In addition, he recommended that encapsulated cell lines or genetically engineered cells should be used for neural transplantation instead of fetal tissue. If cell transplantation was to become a routine clinical procedure in the future, then more donors for the controversial material from aborted human fetuses would be needed and the source of human material itself would be unavailable in some countries (Lindvall 1991).

Ten years later, stem cells seemed to provide a means to settle at least some ethical concerns, although they were also disputed as problematic because of their embryological origin. However, in the eyes of some ethicists, lawyers, and scientists, "the promise of stem cell research for millions of patients" and the different characteristics of stem cells making them less connected to abortion justified their use for therapeutic purposes. In line with the tradition of neural transplantation, "medical progress" was offered as a multi-tool offering at the same time therapeutic hope and a technological fix for ethical concerns (Annas et al. 1999; Merkel et al. 2007, p. 115).

Despite the argument of research offering hope, the problem of the justification of clinical trials in patients remained. It was (and still is (Hess 2012)) disputed as to which patients and under which risk-benefit ratios experimental intracerebral interventions are morally acceptable. Groups such as the "Network of European CNS Transplantation and Restoration," which besides technical aspects also addressed ethical issues (Hoffer and Olson 1991), formulated protocols for operation procedures and recommendations for trial protocols and patient selection. Major moral issues revolved around the irreversibility or only the partial reversibility of neural grafting and gene therapy in the brain. In addition, the question was under discussion of whether sham surgery was acceptable or not, because among other arguments, it was disputed as to whether proper informed consent for sham surgery was possible at all (Merkel et al. 2007, pp. 83–91). Finally, the debate regarding trials in patients was influenced by fears and a lack of knowledge of possible personality changes that could occur as a result of the procedure.

Although Merkel et al. retrospectively regarded the debate about personality changes as "erroneously brought up" (Merkel et al. 2007, p. 104), neural transplantation and gene therapies of the brain had been accompanied by a differentiated philosophical debate about possible changes of personality and identity (Gillon 1996; Northoff 1996; Burd et al. 1998), which has been repeated in the context of stem cell transplantation (Grisolia 2002; Goldstein 2013). Based on thought experiments by researchers such as Derek Parfit among others Georg Northoff discussed the arguments of proponents and opponents of the idea that brain transplantation altered the personal identity of a patient (Goldstein 2011). He concluded that both used the same arguments but with "different underlying presuppositions" regarding the meaning of "identity," the definition of brain identity, and the relationship "between mental states, psychological functions, and neurophysiological properties." Therefore, he urged for neurophysiological, clinical, and philosophical evidence to enrich the discussion (Northoff 1996).

Without directly referring to Northoff, Rehcrona provided some of the needed evidence in a critical review published in 1997 on the status of brain transplantation. Rehcrona evaluated the results of human clinical trials that had been performed up until that year. In the review, he considered the risk of a transfer of personality among other aspects, which he "assumed to be nonexistent," because only small tissue parts, limited amounts of material, and above all undifferentiated embryological tissue are transplanted. If personality changes did occur, then he tended to attribute them to the cure of a disease, which had altered a patient's personality before, and if that was the case then he regarded these changes as "goals of the treatment" rather than a "complicating side effect" (Rehncrona 1997, p. 8). However, he took a different stance regarding psychiatric side effects, which had been reported in some studies. In particular, he stated that after open microneurosurgical grafting (different from stereotactical methods), a considerable number of patients suffered from "frontal lobe syndrome," which by definition is characterized by severe changes of personality. Regarding a possible use of neural grafting to treat psychiatric patients, he directly warned against "premature, human experiments with brain tissue grafting . . . to treat psychiatric disorders" and requested

solid scientific data and methods to reduce these associated risks (Rehncrona 1997). After all, the issue of human identity in close connection to neurocentric concepts of cognition and personhood seems to be one of the hinges on which the ethical debate about alterations – be it grafts of genes – of the very "wetware" of our minds is now swinging (DeGrazia 2005).

Concluding Remarks

Rehncrona's paper seems to paradigmatically reflect the discursive field that has shaped the discussion about neural transplantation during the 1990s and stem cell transplantation into the brain during the 2000s. Hope, risk perception, opaque fears, and personal identity were issues directly linked to the manipulation of patients' brains. The source of the material to be transferred (fetal tissue, stem cells) was the associated or even preceding practical problem that brought with it moral implications.

Similar to this interpretation, the analysis by Moreira and Palladino differentiated the regimes of "hope" and "truth" as two crucial and interdependent tropes of the discourse in the 1990s and early 2000s about treating patients suffering from Parkinson's disease with the help of neurotransplantation. According to their view, the regime of hope justifies experimental and therapeutic endeavors with the promise of what could be, while the regime of truth rather "entails an investment in what is positively known" (Moreira and Palladino 2005, p. 67). While the regime of hope considers the patient as someone waiting for new solutions, the regime of truth sees the patient as someone who evaluates approaches according to risks, effectiveness, and costs. In the debate about neuroimplantation, these regimes rely on each other in "mutual parasitism" (Moreira and Palladino 2005, p. 73), because hope in what could be can only rely on what is positively known so far, and further gains of knowledge rely on hope for the better, because "truth" requires future experiments triggered by hope as a resource of evidence. A point of reference for both regimes in the neuroimplantation debate is seen by Moreira and Palladino in the "self," which shall either be recovered by the therapies in question under the regime of truth or relaunched as a "new self" emerging from neurophysiological dynamics under the regime of hope (Moreira and Palladino 2005, p. 74).

Robert White's experiments of the late 1960s did not offer much hope, and their truth in the form of medical value was questioned by medical scientists (Freed 2000). Thus, his sensational reports were subsequently only discussed in news magazines as frightening Frankenstein stories after the 1980s. The other forms of neural transplantation discussed here, however, were able to bring the regimes of hope and truth together at the same time. While some moral issues, such as the use of fetal material, could be seemingly fixed technologically (the technological fix is problematic under some perspectives, because the technology per se is not discussed (Fangerau 2011)), other moral issues revolving around hope and truth, risk and fear, and prospects for personal identity could not be fixed and are debated in other chapters of this handbook.

Cross-References

▶ Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges
▶ Ethics in Neurosurgery
▶ Impact of Brain Interventions on Personal Identity
▶ Neuroethics and Identity
▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives
▶ Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia

References

Anderson, W. F. (1990). Editorial. September 14, 1990: The beginning. *Human Gene Therapy, 1*(4), 371–372.

Annas, G. J., Caplan, A., et al. (1999). Stem cell politics, ethics and medical progress. *Nature Medicine, 5*(12), 1339–1341.

Anonymous. (1890). Successful brain grafting. *Science, 16*(392), 78–79.

Anonymous. (1968, August 14). Controversy surprises Barnard. *Times, Overseas News*, 3.

Anonymous. (1976). Kopf mit Kraftpaket. *Der Spiegel, 44*, 225.

Anonymous. (1982, November 22). Transplant hopes for brain tissue. *Times*, *Home News*, 2.

Anonymous. (1983, August 20). Work in rats. *The Economist, 7303*, 85–86.

Anonymous. (2001, April 6). Frankenstein fears after head transplant. *BBC News Online*. Retrieved March 3, 2013, from http://news.bbc.co.uk/2/hi/health/1263758.stm

Backlund, E. O., Granberg, P. O., et al. (1985). Transplantation of adrenal medullary tissue to striatum in parkinsonism. First clinical trials. *Journal of Neurosurgery, 62*(2), 169–173.

Beardsley, T. (1990). Aborted Research. Ideology seems to have put some medical advances on hold. *Scientific American, 262*(2), 16.

Berg, T. (2006). Human brain cells in animal brains: Philosophical and moral considerations. *The National Catholic Bioethics Quarterly, 6*(1), 89–107.

Bishop, J. M. (1982). Oncogenes. *Scientific American, 246*, 80–92.

Björklund, A., & Stenevi, U. (1985a). Intracerebral neural grafting: A historical perspective. In A. Björklund & U. Stenevi (Eds.), *Neural grafting in the mammalian CNS* (pp. 3–14). Amsterdam: Elsevier.

Björklund, A., & Stenevi, U. (Eds.). (1985b). *Neural grafting in the mammalian CNS*. Amsterdam: Elsevier.

Boer, G. J. (1994). Ethical guidelines for the use of human embryonic or fetal tissue for experimental and clinical neurotransplantation and research. *Journal of Neurology, 242*(1), 1–13.

Boer, G. J. (1999). Ethical issues in neurografting of human embryonic cells. *Theoretical Medicine and Bioethics, 20*(5), 461–475.

Borges, L. F. (1988). Historical development of neural transplantation. *Applied Neurophysiology, 51*(6), 265–277.

Boyer, K. L., & Bakay, R. A. (1995). The history, theory, and present status of brain transplantation. *Neurosurgery Clinics of North America, 6*(1), 113–125.

Burd, L., Gregory, J. M., et al. (1998). The brain-mind quiddity: Ethical issues in the use of human brain tissue for therapeutic and scientific purposes. *Journal of Medical Ethics, 24*(2), 118–122.

Carlson, G. A., Borchelt, D. R., Dake, A., Turner, S., Danielson, V., Coffin, J. D., Eckman, C., Meiners, J., Nilsen, S. P., Younkin, S. G., & Hsiao, K. K. (1997). Genetic modification of the phenotypes produced by amyloid precursor protein overexpression in transgenic mice. *Human Molecular Genetics, 6*(11), 1951–1959.

Clinton, W. J. (1994). Memorandum of January 22, 1993. *Code of federal regulations; Title 3, The President* 1993 compilation and parts 100 to 102, 724.

Coutts, M. C. (1993). Scope Note 21. Fetal tissue research. *Kennedy Institute of Ethics Journal, 3*(1), 81–101.

Das, G. D. (Ed.). (1986). *Neural transplantation and regeneration*. Berlin/Heidelberg/New York: Springer.

Das, G. D. (1990). Neural transplantation: An historical perspective. *Neuroscience and Biobehavioral Reviews, 14*(4), 389–401.

DeGrazia, D. (2005). *Human identity and bioethics*. New York: Cambridge University Press.

Dunnett, S. B. (Ed.). (1992). *Neural transplantation: A practical approach*. Oxford: IRL Press.

Fangerau, H. (2011). Brain, mind and regenerative medicine: Ethical uncertainties and the paradox of their technical fix. In H. Fangerau, J. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 15–30). Bielefeld: Transcript.

Freed, W. J. (2000). *Neural transplantation. An introduction*. Cambridge, MA: MIT Press.

Fujimura, J. H. (1992). Crafting science: Standardized packages, boundary objects and 'translation'. In A. Pickering (Ed.), *Science as practice and culture* (pp. 168–211). Chicago: The University of Chicago Press.

Gash, D. M. (1984). Neural transplants in mammals: A historical overview. In J. R. Sladek & D. M. Gash (Eds.), *Neural transplants: Development and function* (pp. 1–12). New York: Plenum Press.

Gash, D. M., & Sladek, J. R., Jr. (Eds.). (1988). *Transplantation into the mammalian CNS: Based on the Schmitt Symposium on Transplantation into the Mammalian Central Nervous System, held June 30 – July 3, 1987 in Rochester, NY*. Amsterdam: Elsevier.

Gash, D. M., Collier, T. J., et al. (1985). Neural transplantation: A review of recent developments and potential applications to the aged brain. *Neurobiology of Aging, 6*(2), 131–150.

Gillon, R. (1996). Brain transplantation, personal identity and medical ethics. *Journal of Medical Ethics, 22*(3), 131–132.

Goldstein, J. (2011). Parfit's concept of personal identity and its implications for intercerebral stem cell transplants. In H. Fangerau, J. Fegert, & T. Trapp (Eds.), *Implanted minds. The neuroethics of intracerebral stem cell transplantation and deep brain stimulation* (pp. 15–30). Bielefeld: Transcript.

Goldstein, J. (2013). Personale Identität und intrazerebrale Stammzelltransplantationen. Diss. Med. Fac. Ulm University.

Greene, M., Schill, K., et al. (2005). Ethics: Moral issues of human-non-human primate neural grafting. *Science, 309*(5733), 385–386.

Grisolia, J. S. (2002). CNS stem cell transplantation: Clinical and ethical perspectives. *Brain Research Bulletin, 57*(6), 823–826.

Hess, P. (2012). Intracranial stem cell-based transplantation: Reconsidering the ethics of phase 1 clinical trials in light of irreversible interventions in the brain. *AJOB Neuroscience, 3*(2), 3–13.

Hoffer, B. J., & Olson, L. (1991). Ethical issues in brain-cell transplantation. *Trends in Neurosciences, 14*(8), 384–388.

J. D. (1891, March 1). Brain transplantation. *New York Times*, *18*.

Jones, D. G. (1991). Fetal neural transplantation: Placing the ethical debate within the context of society's use of human material. *Bioethics, 5*(1), 23–43.

Jones, E. G., & Mendell, L. M. (1999). Assessing the decade of the brain. *Science, 284*(5415), 739.

Lindvall, O. (1991). Prospects of transplantation in human neurodegenerative diseases. *Trends in Neurosciences, 14*(8), 376–384.

Lindvall, O., Björklund, A., et al. (1991). *Intracerebral transplantation in movement disorders: Experimental basis and clinical experiences* (Proceedings of the 20th Fernström symposium, Lund, Sweden, 1990). Amsterdam: Elsevier.

Low, W. C., Lewis, P. R., et al. (1982). Function recovery following neural transplantation of embryonic septal nuclei in adult rats with septohippocampal lesions. *Nature, 300*(5889), 260–262.

Mahowald, M. B., Areen, J., et al. (1987). Transplantation of neural tissue from fetuses. *Science, 235*(4794), 1307–1308.

Martin, M., & Fangerau, H. (2006). Die Bedeutung unterschiedlicher Textsorten für die Repräsentation von Wissensverschiebungen in der Medizingeschichte. In J. Vögele, H. Fangerau, & T. Noack (Eds.), *Geschichte der Medizin – Geschichte in der Medizin – Forschungsperspektiven* (pp. 153–162). Münster: Lit Verlag.

McCullagh, P. (1988). Fetal brain transplantation – The scope of the ethical issue. *Ethics & Medicine: A Christian Perspective on Issues in Bioethics, 4*(3), 37–39.

Merkel, R., Boer, G. J., et al. (Eds.). (2007). *Intervening in the brain: Changing psyche and society* (Ethics of science and technology assessment). Berlin/New York: Springer.

Moreira, T., & Palladino, P. (2005). Between truth and hope: On Parkinson's disease, neurotransplantation and the production of the 'self'. *History of the Human Sciences, 18*(3), 55–82.

Morrison, K. E. (1987). Brain transplantation – Still fantasy? Discussion paper. *Journal of the Royal Society of Medicine, 80*(7), 441–444.

Moss, M. B., & Rosene, D. L. (1985). Neural transplantation – A panacea. *Neurobiology of Aging, 6*(2), 168–169.

Northoff, G. (1996). Do brain tissue transplants alter personal identity? Inadequacies of some "standard" arguments. *Journal of Medical Ethics, 22*(3), 174–180.

Paul, N. W. (2003). Risiko, Sicherheit, Nutzen und Strategien zur Implementierung von Gentherapien 1980–2000. In A. Labisch & N. W. Paul (Eds.), *Historizität: Erfahrung und Handeln, Geschichte und Medizin* (pp. 201–210). Stuttgart: Steiner.

Rehncrona, S. (1997). A critical review of the current status and possible developments in brain transplantation. In F. Cohadon, V. V. Dolenc, J. L. Antunes, et al. (Eds.), *Advances and technical standards in neurosurgery* (Vol. 23, pp. 3–46). Vienna: Springer.

Rogers, J., Zornetzer, S. F., et al. (1985). Therapeutic applications of neural transplant technology. *Neurobiology of Aging, 6*(2), 169–172.

Sanberg, P. R. (1990). Students' views on fetal neural tissue transplantation. *Lancet, 335*(8705), 1594.

Thompson, W. G. (1890). Successful brain grafting. *New York Medical Journal, 51*, 701–702.

U.S. Congress. Office of Technology Assessment (Ed.). (1990). *Neural grafting: Repairing the brain and spinal cord, OTA-BA-462*. Washington, DC: U.S., Government Printing Office.

Wade, N. (2005, July 15). Ethicists offer advice for testing human brain cells in primates. *New York Times, A12*.

Wallace, R. B., & Das, G. D. (Eds.). (1983). *Neural tissue transplantation research*. New York: Springer.

White, R. J. (1985). Brain transplantation. *Surgical Neurology, 23*(4), 449.

White, R. J. (1992). Fetal brain transplantation: Questionable human experiment. *America (NY), 167*(17), 421–422.

Widmann, A. (2000, September 1). Der Quartals-Frankenstein. *Berliner Zeitung*. Retrieved March 3, 2013, from http://www.berliner-zeitung.de/archiv/robert-white-will-koepfe-von-menschen-verpflanzen–bei-einem-vortrag-in-dresden-wurde-deutlich–dass-er-vor-allem-in-den-koepfen-etwas-bewegt-der-quartals-frankenstein,10810590,9829568.html

Wilberger, J. E. (1983). Transplantation of central nervous tissue. *Neurosurgery, 13*(1), 90–94.

Gene Therapy and the Brain

54

Christian Lenk

Contents

Abstract

The chapter examines the implications of gene therapy in the human brain from the ethical point of view. In the first part, a number of principles from research ethics are discussed for the field of the neurosciences. While there exists a consensus in the area of medical ethics as a whole, what ethical principles are essential, this is less clear in the specialized field of neuroethics. Therefore, the presentation and discussion of the seven principles has the aim to clarify the normative foundations of the following ethical evaluation of gene therapy in the brain. In the second part, selected studies with clinical applications are presented from the literature and general risks and burdens for participating patients are identified. The literature review shows a participation of vulnerable persons, and the use of sham surgery and randomization for the blinding of studies. The situation of patients with a chronic dementia disease is addressed, and conclusions are derived under an ethical perspective. In this chapter's final section, a short overview is given about the current development of gene therapy from the ethical point of view, and which advances and setbacks seem to be important for the assessment of the research field. Gene therapy for neurological diseases

C. Lenk
Ulm University, Ulm, Germany
e-mail: christian.lenk@uni-ulm.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_128,

shows some specific risks and burdens for participating patients. These have to be addressed carefully and systematically to realize adequate research study designs, to avoid misunderstandings on the side of the patients and study participants, and to enable an optimal risk monitoring and risk control in the clinical testing and use of this new therapeutic approach.

Introduction and Basic Ethical Considerations

Before analyzing the field of gene therapy in the brain and the connected ethical questions, from a medical ethics' point of view, it is preferable first to clarify the underlying ethical principles which form the basis of our normative work. This task is not so difficult for the whole field of medical ethics, because there is a kind of consensus that the four-principle approach of Beauchamp and Childress can serve as a starting point for ethical analysis in the medical field. However, despite the fact that the four-principle-approach can give some orientation, it is also true that it is in most cases not sufficient to enable a fully comprehensive understanding of ethical problems. And the normative foundations seem to be even more unclear in the case of neuroethics, because less publications have focused on this issue and at the present point of time, no acceptable consensus has appeared [Insert reference to theoretical articles in the handbook?], what norms and values are decisive for the ethical analysis of neurological problems. Therefore, I want to create something like an ad hoc approach for the analysis in this chapter and discuss some principles of research ethics and neuroethics which were prominent in the previous discussion. A number of these principles are also described in international conventions and declarations, what shows their relevance over and above the narrow field of interest in this chapter. The idea is not to develop at this place a new ethical approach to neuroethics but rather to clarify the normative issues as far as possible before starting to discuss the ethical problems connected with gene therapy and the brain.

The fact that gene therapy is at the present point of time an experimental approach has also some influence on our perception of the field. In the application of an already introduced therapy, there are clear expectations, which proportion of the patients can profit from a specific therapy. Ideally, there are also a number of well-defined criteria which allow identifying those persons of a patient group who will especially profit from this therapy. These expectations are derived from former practical experiences, already conducted clinical studies, and so on. However, the focus in experimental studies is different, due to the fact that there is at the starting point of a clinical trial, no confirmed knowledge regarding the effects of this new therapy. Therefore, the focus in research ethics lies far more on the responsible limitation of risk than on possible benefits for the patients. Strategies of risk limitation comprise, among others, preclinical and animal studies, pharmacological analysis and evaluation, close medical and psychological monitoring of patients during the study, the appropriate choice of the participating patient group, and a rather conservative risk-benefit-analysis. Researchers, physicians, and patients,

especially in the case of dangerous, severe, and chronic diseases, are often enthusiastic when confronted with new research approaches. Research ethics faces here the difficult task to transfer previous experiences and sometimes difficult lessons from research to the clinical practice.

Some Normative Principles for Neuroethics

1. *Every medical intervention must respect the autonomy and dignity of patients.*

   It can count as one of the premises of modern biomedical research, that it has to be in accordance with the patients' autonomy and dignity or that it is otherwise not acceptable. The patients' autonomy has to be guaranteed by the patient's informed consent or the assent of her or his relatives or legal representatives. Despite the mere self-determination of patients through the principle of autonomy (which is materialized in the written consent of the patient), human dignity can be seen as a even wider and more comprehensive ethical demand which (in the Kantian tradition) pronounces a person's intrinsic value and aims to exclude the instrumentalization of persons, i.e., to use somebody "merely as a mean." Therefore, it can also be found in Art. (1) of the CoE Convention on Human Rights and Biomedicine.[1]
2. *The patient's original identity should be preserved or restored in the course of a neurological intervention as far as this is possible.*

   The identity of persons was without doubt one of the central concerns of the previous discussion on neuroethics. Like other medical interventions which aim on the restoration of an original (healthy) state, for example, the treatment of demented patients in neurology aims also on a kind of restoration of an "original" mental state in the sense of "not influenced by disease-related dysfunctions." However, it has to be considered that personal identity is not a static, but a dynamic concept and in the case of neurological diseases, that the personal identity can also be influenced by nonreversible pathological developments. However, from my point of view, it is a clear implication of neurological *therapy*, that it is in general not the aim of such a therapy to create a person with a new or changed identity but to preserve and restore somebody's original identity as far as this is possible (in the sense of a *restitutio ad integrum*). These short descriptions already show that there is an urgent need for more theoretical considerations in neuroethics concerning this important but very complex issue.
3. *Research interventions in competent persons are from the ethical point of view preferable in comparison to interventions in noncompetent persons. Research interventions in noncompetent persons are only acceptable with a high probability to realize a significant health benefit for the concerned person.*

[1]"Parties to this Convention shall protect the dignity and identity of all human beings and guarantee everyone, without discrimination, respect for their integrity and other rights and fundamental freedoms with regard to the application of biology and medicine. Each Party shall take in its internal law the necessary measures to give effect to the provisions of this Convention."

Accordingly, the Council of Europe's Convention on Human Rights and Biomedicine, foresees in Art. (17), 1 that:

> Research on a person without the capacity to consent [. . .] may be undertaken only if all the following conditions are met:[. . .]ii. the results of the research have the potential to produce real and direct benefit to his or her health;iii. research of comparable effectiveness cannot be carried out on individuals capable of giving consent; [. . .]v. the person concerned does not object.

Following the Convention, an exception to this rule is only given, when "the research entails only minimal risk and minimal burden for the individual" and the research contributes to the scientific understanding of the concerned persons' or comparable persons' condition. The Convention's additional protocol on biomedical research concretizes in Art. (17), 2 that a minimal burden is only a "temporary and very slight [discomfort]." In another document, the EU Ad Hoc Group (2008) names for the pediatric population a quality-of-life assessment or blood pressure monitoring as examples for a "minimal risk."

4. *In medical research in general and experimental settings in particular, the potential therapeutic benefits of neurological interventions have to be weighed very carefully against possible risks and adverse events.*

   May be, one could call this clause also the "proportionality principle" of research ethics.[2] This principle can be found in most documents on the correct ethical and legal conduct of research in humans. Art. (16) of the Declaration of Helsinki foresees that "Medical research involving human subjects may only be conducted if the importance of the objective outweighs the risks and burdens to the research subjects." Comparably, the Oviedo Convention from the Council of Europe demands in Art. (16): "Research on a person may only be undertaken if all the following conditions are met: [. . .] (ii) the risks which may be incurred by that person are not disproportionate to the potential benefits of the research; [. . .]" Given that the exact possible benefits of gene therapy in the brain are at the present point of time still unclear, every effort should be undertaken to properly assess and minimize possible risks for the participating patients.

5. *Reversible interventions are preferable to irreversible interventions.*

   A concern which is frequently articulated in the context of innovative and experimental methods is that irreversible interventions should only be undertaken at a point of time when reversible interventions have shown not to be effective. The idea behind this principle is that reversible interventions could ideally be made undone, but that irreversible interventions will accompany the patient far into the future. Therefore, the reversible intervention should be preferred in such a situation. For example, one would prefer in a specific patient group as a first-line treatment an established medication before implanting a medical device for deep brain stimulation. However, also electrodes for deep brain stimulation are rather reversible like older forms of neurosurgery which

[2]For a comprehensive and current overview on proportionality in research ethics, cf. Hermerén (2012).

were definitely irreversible. But one has also to consider that the human body has no mere modular structure and that also the extraction of medical devices from the body can bear a significant burden and risk for the concerned patient.

6. *In neurosurgical interventions, not only the direct and short-time consequences should be documented, but there should also be a systematic plan to document and evaluate the long-time effects on the patients' health, personality, and quality of life.*

   It is a known claim in clinical studies which aim on the approval of a new drug that the application of this drug should also be further observed in clinical practice. For example, in an approval study, the patient population is often different from the patients who will receive the drug finally in medical practice. Additionally, it is not known whether the first studies with a drug are sufficient to document all possible side effects. For example, a long-term study revealed in 2004 that the analgesic Vioxx had unacceptable side effects for the patients. Neurosurgical and genetic interventions into the brain should also be reassessed by such long-term observational studies.

7. *The introduction of experimental approaches should take place very carefully, only in small steps, and only after the sufficient preparation by preclinical and animal studies.*

   Insofar gene therapy in the human brain is a new and experimental therapeutic approach with no known predecessors, research in human beings has to be prepared by animal studies, and it has to be started only in small numbers in specially selected patient populations. On the one hand, patients should be able to give their informed consent and should already show manifest signs of disease; on the other hand, experimental therapies with critically ill patients in the last stages of their life have to be seen as problematic. It seems to be more appropriate for experimental tests of fully new therapeutic approaches to start clinical studies with patients in a stable health state and a maximum of safety measures (careful preclinical and animal tests, low doses for the first clinical studies, close medical and psychological monitoring during tests, and so on).

Genetic Interventions and Connected Risks in Clinical Research Practice

A general ethical assessment of gene therapy in the brain seems to be difficult at the present point of time due to the structure, location, and organization of research. From a short literature review (Berry and Foltynie 2011; Kaplitt et al. 2007; LeWitt et al. 2011), it follows that such studies are currently only organized with rather small patient groups in early study phases which deal with a general proof of principle or the safety and tolerability of the respective therapeutic approach. The different approaches – in relation to the kind of virus for gene transfer and the kind of inserted genes – seem to be rather heterogeneous and diverse. It follows that a comprehensive and concluding discussion of the therapeutic benefit of gene therapy in the brain cannot be done at the moment. But this means also that

a final risk-benefit-analysis, which must always be an essential element of an ethical analysis in research ethics, is difficult at the present point of time. The report of therapeutic effects is in the end – despite the publication of research progresses – rather anecdotal and does not reach the level of reliability, for example, of large clinical drug trials.

In place of an overview of the different therapeutic approaches to a number of disease entities, I will focus in this chapter on the area of gene therapy in the brain which seems to be at the present point of time the most advanced, i.e., therapeutic approaches in Parkinson's disease (PD). Therefore, the above named three studies were selected from the last 5 years which were published in high-profile scientific journals following the hypotheses that ethical questions concerning gene therapy in the brain can be probably identified best in the most advanced cases of a research paradigm. However, it cannot be concluded in any case that it is possible to transfer the findings from this research field to other applications of gene therapy in the brain.

The application of gene therapy in the case of the human brain must produce a number of ethical questions and different kinds of risks due to the combination of different forms of invasive interventions.

Firstly, the current research approaches of gene therapy in the brain need a surgical intervention to bring the therapeutic substances to the place where they are intended to have their therapeutic effect, i.e., into the areas of the brain which are influenced by the pathological processes of the disease. However, such a therapeutic approach then produces an own kind of risk with the one-time or repeated injection of the therapeutic agent into the brain or the implantation of a cannula for the supply of the treated brain areas (with the opening of the skull, the irritation and swelling of tissue, possible infection, and so on).

Secondly, there is the known genuine risk of a gene therapy treatment. In the historic case of a volunteer who died in 1999 as a test person in a Phase 1 dose escalation study concerning ornithine transcarbamylase (OTC) deficiency, the 18-year-old developed a fatal immune response due to the high dose of adenoviral vectors (Kimmelman 2008, p. 242). Immune reactions of study participants in the case of adeno-associated virus (AAV) vectors are also cited in a later gene therapy study (Kaplitt et al. 2007, p. 2097). In 2003, three of the children, who suffered from X-linked severe combined immune deficiency (X-SCID) and were treated with a new gene therapeutic approach at the French Necker Hospital, developed T-cell leukemia (Wood et al. 2006; Zarzer 2006).[3] Further research showed that the leukemia was provoked by the insertion of the changed gene material. The authors of the named articles showed also that this danger could have been identified before the clinical trials when animal tests would have been longer than the 6 months before the actual trial (Wood et al. 2006). The review article of Berry and Foltynie (2011) additionally names antibody formation and cerebellar toxicity in animal

[3]In 2008, Kimmelman reports that this number has later increased to five children with leukemia. (Kimmelman 2008, p. 239).

Table 54.1 Combination of different types of risks in the case of genetic interventions into the brain

Kind of intervention / effect	Possible associated risks
1. Neurosurgical interventions for the transfer of the therapeutic agent into the brain	Irritation and swelling of tissue, possible infection, etc.
2. Genetic intervention (gene expression and gene transport by viruses)	Immune reactions, antibody formation, cerebellar toxicity, cancer, disturbance of dopamine concentration or of other substances
3. Long-term neurological or psychological impact of (2)	Yet not described for genetic interventions into the brain, possible change of personality (see for example the debate on deep brain stimulation)

models (Berry and Foltynie (2011), p. 180), the disturbance of dopamine concentration in the brain (Berry and Foltynie (2011), p. 183), and an "indiscriminate gene expression across all neuronal cells (and/or glial cells)" (Berry and Foltynie (2011), p. 183) as possible adverse events of gene therapy in the case of Parkinson's disease. Kimmelman (2008, p. 239) concludes:

> What distinguishes the risks of somatic gene transfer trials from those for conventional drugs is not so much the level of risk [. . .] but rather their level of complexity and of uncertainty. [. . .] Furthermore, even though numerous trials involving retroviral vectors have been carried out, there is no widely accepted system for quantifying the risks of insertional mutagenesis.

Thirdly, in the case of cerebral interventions and manipulation, a possible change of personality of the treated person is constantly discussed and would conflict with our principle (2) which was described at the start of this article ("*A patient's original identity should be preserved or restored in the course of a neurological intervention as far as this is possible*.") Changes of personality were described for the "classic" neurosurgical interventions like lobotomy in the 1940s and 1950s as well as for the current approaches to treat motoric and psychiatric symptoms, for example, of Parkinson's Disease with deep brain stimulation. It has to be considered that there can be huge differences in the extent of personality changes from the strong and irreversible reduction of personality traits (in the case of lobotomy) to relatively slight and transient hypomanic episodes (in the case of deep brain stimulation). However, it cannot be excluded that changes in personality can also occur in the case of genetic interventions, especially when these – as described in the last paragraph – will influence the biochemical homeostasis of the brain (Table 54.1).

Examples from Current Studies

From the diseases of the brain, Parkinson's disease (PD) seems to be the disease entity which stands at the moment most directly in the focus of gene therapy. The disease manifests relatively lately in life and is in general associated with

movement disorder and tremor, but also with cognitive, emotional, and vegetative changes (Riess et al., A-2739). Pathological research showed that in the brain tissue of concerned patients, the cell death of specific neurons leads to a lack of dopamine in the striatum (a part of the forebrain) which is then (according to this model of disease) seen as the cause of the movement disorder. This part of the brain is in general associated with the so-called executive brain functions which coordinate emotion, cognition, and movement. The frequency of PD is dependent from age and increases from 1.4 % in the group of the 55 years old to 2.0 % at an age of 65 and 3.4 % in the group of the 75-year-old persons (Riess at al., A-2742). Although hereditary forms of PD are known, the majority of cases is multifactorial or polygenic (Riess et al., A-2746). The idea in a number of PD studies based on gene therapy is in general to insert genetic material into the brain cells to increase the dopamine production and therefore to compensate the tissue's lost ability to produce this neurotransmitter.

From medical ethics' point of view, patients with dementia are a group of "vulnerable patients" and their inclusion in medical research needs specific attention and consideration. However, not because of the disease entity itself patients with PD are seen as vulnerable, but because of the possible disease impact on patient autonomy, there may be impaired ability for self-determination, the competence to assess the consequences of possible study participation, and maybe a diminished ability to give informed consent. Therefore, persons in an early phase of the disease may act fully competent while patients in a later phase of the disease may have lost this ability (cf. also Rosenstein and Miller 2008, figure 41.1: 439). In addition to the ethical considerations concerning the treatment of these patients, many national legislations demand the designation of a legal guardian for patients with reduced competence. The Declaration of Helsinki and the Council of Europe's Declaration on Human Rights and Biomedicine both have own paragraphs on medical research with patients who are not fully competent to give informed consent. For the participation of noncompetent participants, in addition to the approval of the legal guardian, the Declaration of Helsinki demands in Art. (28) that:

> These individuals must not be included in a research study that has no likelihood of benefit for them unless it is intended to promote the health of the group represented by the potential subject, the research cannot instead be performed with persons capable of providing informed consent, and the research entails only minimal risk and minimal burden.

In the case of experimental, risky, and invasive treatments (like in the case of gene therapy in the human brain), there is often the consideration that only patients should be included into the study which did not profit from other (for example, pharmaceutical) therapeutic approaches. The rationale behind this procedure is that those who obviously have no further chance of healing in conventional therapeutic regimes should receive a "last chance" to participate in a therapeutic study, even when the probability of a successful outcome might be rather small. But this can lead to the effect that only severely ill (and in the case of PD, noncompetent) patients are included into such a study. However, from my point of view, such an approach should be limited to less experimental therapeutic approaches and, like it

was described above, in the case of gene therapy, only patients in earlier phases of the disease should be included into such studies.

Current research approaches of gene therapy in the case of PD mostly base on the creation of viruses which have the task to bring gene particles into the concerned body region (the brain), where the changed genetic information should get into the defective cells or tissue.

One of the selected studies was a sham-surgery controlled, randomized trial (LeWitt et al. 2011). Sham-surgery control is a procedure where a part of the participating patients does not receive the injection of the therapeutic agent into the brain tissue, but only a placebo. To mask this fact and to enable the so-called blinding of the study, these patients receive the same surgical intervention, but no injection into the brain tissue (although, for example, the noise of the pumps is simulated, and so on; LeWitt et al. 2011, p. 310). Placebo-controlled trials are rather recommended for slighter diseases, where the symptom assessment is also clearly influenced by subjective factors. Due to the severity of the symptoms of PD, it is questionable whether it should be used in the case of this disease.

Sham surgery produces also new risks and burdens for the participating patient group without an individual benefit. The most commonly accepted document of research ethics, the Declaration of Helsinki, does not address the conduct of sham surgery for blinding purposes. However, it is generally presumed that the blinding of a study with the help of placebo (in drug testing) itself does not harm the patient. This is also the reason why severe disease entities are normally not accepted for placebo studies, because the omission of therapy would otherwise provoke further disease progress and therefore a study-related harm of the patient. However, the case is obviously different in case of sham surgery, where the patient carries in any case the burden and possible health risk associated with sham surgery. From my point of view, the possible scientific value does not justify the connected burden and risk for PD patients. Like it was described above, the method of gene transfer bears itself remarkable health risks, and methodological enthusiasm should be rather invested into the development of safer gene transfer than into avant-garde methods of study blinding.

Conclusions

This chapter presented a general overview about the previous developments of gene therapy from the point of view of patient safety and medical ethics. This overview was combined with an analysis of current studies with PD patients of gene therapy in the brain. For both areas, sufficient examples and well-documented studies could be identified for the following conclusions from medical ethics. However, it has to be seen that in comparison to other (i.e., pharmacological) treatment approaches of the disease, the present descriptions and studies of gene therapy in the brain are rather experimental and based on rather small patient groups. Statements on risks and benefits of gene therapy in the brain must therefore remain tentative at the present point of time.

The historical development of gene therapy shows impressive progresses and successes. In a way, it is astonishing that the insertion and expression of therapeutically manipulated gene material into the cells of the human body is possible. However, our analysis showed that the development of gene therapy is accompanied by hard and maybe disastrous setbacks, which caused the death or severe disease of participating patients, who belonged to vulnerable groups like adolescents and children. The medical and ethical discussion after these severe adverse events showed that they would have been in principle avoidable by a more careful study design (i.e., avoidance of dose escalation and longer and more comprehensive animal and preclinical studies). However, because the researchers were not able to foresee the existing problems, these adverse events could not be prevented.

The normative (ethical and legal) evaluation mechanisms and principles have probably to develop together with research in gene therapy and neurology. This is at least a conclusion one can draw from other fields of medical and research ethics, because in most of the research fields, it cannot correctly be anticipated what direction practical research will take and what kind of burdens and risks patients may face with new research approaches and study designs. The existing framework for ethical analysis and evaluation has therefore to be strengthened and systematized to achieve a better consensus about the relevant ethical criteria. In comparison to other fields of medical ethics, the ethical evaluation of gene therapy in the brain as an interdisciplinary area between neurology and gene therapy has at the present point of time to resort to a kind of ad hoc compilation of ethical principles from the previous discussion.

The ethical analysis of gene therapy in the brain in the case of PD patients identified a number of potential risks and burdens for study participants. Firstly, it should be further cleared and discussed, which PD patients should be included into clinical studies. Given that gene therapy comprises more than “minimal risk,” the participants should be able to decide for themselves to give the consent for study participation. Secondly, gene therapy can produce totally different kind of risks (for example, immune responses, increased risk of neoplasia, etc.) which strongly seem to differ between different studies. It follows that there must be an extremely careful case-to-case evaluation of gene therapy studies. Finally, the evaluation of innovative research projects has to a certain degree also to consider the context of these projects, like it is described by Kimmelman (2008, p. 240):

> Many early phase gene-transfer trials bring together a potent mixture of desperately ill research subjects, ambitious (and sometimes financially invested) clinical champions, biotechnology firms, engaged disease advocates and news media. These factors have at times produced a turbulent dynamic in which concepts are rushed into trials, preclinical and clinical results are oversold, research efforts are fragmented and adverse events go underreported.

In principle, research ethics and the existing documents like the Declaration of Helsinki are prepared to deal with these problems. However, a steady and systematic backing of such research projects is necessary not only to realize potential benefits from gene therapy, but to avoid unacceptable risks and burdens for the participating patients.

Cross-References

▶ Biosecurity as a Normative Challenge
▶ Ethical Implications of Cell and Gene Therapy
▶ Ethics of Sham Surgery in Clinical Trials for Neurologic Disease
▶ Human Brain Research and Ethics
▶ Neural Transplantation and Medical Ethics: A Contemporary History of Moral Concerns Regarding Cerebral Stem Cell and Gene Therapy

References

Berry, A. L., & Foltynie, T. (2011). Gene therapy: A viable therapeutic strategy for Parkinson's Disease? *Journal of Neurology, 258*, 179–188.

Council of Europe, Convention on Human Rights and Biomedicine, Oviedo 1997.

EU Ad Hoc Group (for the Development of Implementing Guidelinesfor Directive 2001/20/EC), Ethical Considerations for ClinicalTrials on Medicinal Products Conducted with the Paediatric Population. Brussels 2008. www.ftp.cordis.europa.eu/pub/fp7/docs/ethicalconsiderations-paediatrics_en.pdf. Accessed 20 May 2011.

Hermerén, G. (2012). The principle of proportionality revisited: interpretations and applications. *Medicine, Health Care, and Philosophy, 15*(4), 373–382.

Kaplitt, M. G., Feigin, A., Tang, C., et al. (2007). Safety and tolerability of gene therapy with an adeno-associated virus (AAV) borne *GAD* gene for Parkinson's disease: An open label, phase I trial. *Lancet, 369*, 2097–2105.

Kimmelman, J. (2008). The ethics of human gene transfer. *Nature Reviews. Genetics, 9*, 239–244.

LeWitt, P. A., Rezai, A. R., Leehey, M. A., et al. (2011). AAV2-*GAD* gene therapy for advanced Parkinson's disease: A double-blind, sham-surgery controlled, randomised trial. *Lancet Neurology, 10*, 309–319.

Rosenstein, D. L., & Miller, F. G. (2008). Research involving those at risk for impaired decision-making capacity. In E. J. Emanuel, C. Grady, R. A. Crouch, R. K. Lie, F. G. Miller, & D. Wendler (Eds.), *The Oxford textbook of clinical research ethics* (pp. 437–445). Oxford: Oxford University Press.

Wood, N. B., Bottero, V., Schmidt, M., von Kalle, C., & Verma, I. M. (2006). Therapeutic gene causing lymphoma. Insight into risks posed by corrective gene therapy comes from an immunodeficient mouse model. *Nature, 440*, 1123.

World Medical Association, Declaration of Helsinki, Fortaleza 2013.

Zarzer, B. (2006). TückischeGentherapie? [Treacherous Gene Therapy?] Telepolis, 13.05.2013; www.heise.de/tp/artikel/22/22656/1.html. Accessed 10 Dec 2012.

Section XII

Ethics in Psychiatry

Ethics in Psychiatry

55

Hanfried Helmchen

Contents

Abstract

Good relationships among people have been stabilized by traditional customs and normative rules that have emerged over long time periods and have been acquired during individual development. These customs were originally supported by religious funding; today however they are mainly based on law and social context. Reflecting on customs and morals, on their rational justification, and on analysis of their problem-related impact is the object of ethics. Ethics in psychiatry deals with the reflection on how general moral norms are applied to specific psychiatric situations,[1] especially by the professional physician acting with the mentally ill patient, e.g., respect for the will of the mentally ill with questionable or lost capacity to consent, or risk-benefit-evaluations.

Ethical Principles

Today **welfare** and **will** of the patient are the dominating principles of the ethical evaluation of medical action. Whereas it is timeless Hippocratic tradition that the

[1] Insofar there is no special psychiatric ethics.

H. Helmchen
Department of Psychiatry & Psychotherapy, Charité – University Medicine Berlin, CBF, Berlin, Germany
e-mail: hanfried.helmchen@charite.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_53,

physician should act exclusively for the welfare and in the best interest of the patient, his obligation to respect the autonomy and dignity of the patient has modern roots in the era of enlightenment (Beauchamp and McCullough 1994).

The latter-mentioned principle achieved outstanding significance during the past decades. This is clearly recognizable in the development of the legal doctrine of **informed consent** (Koch et al. 1996). Its sources, among others, are found, on the one hand, in the shifting of modern medicine from acute to long-term treatments, to risky and burdening treatment measures, to replacement medicine, and even to clinical research, all of which require much more participation and thereby more responsibility of the patient, e.g., for adherence in long-term treatments; on the other hand, civil rights, such as the legally protected self-determination and human dignity, will be increasingly recognized as indicated by the UN Convention on Human Rights of Persons with Disabilities (CRPD 2006). The dignity and autonomy of a patient is the more threatened the more he or she is dependent upon nonself-determined actions by which he or she becomes an object, as is found in modern medicine with its widely regulated courses of action. A reflection on this and a change of professional as well as of societal attitudes towards people with mental illness, i.e., to attenuate discrimination and to enhance self-determination, are currently taking place.

One of the old principles, already stated in the Hippocratic Oath, is that of **nil nocere**, i.e., do not harm. This principle emerged in previous times of a wide lack of therapeutic power and of nonproven therapeutic measures. However today, in view of the risks and unwanted side effects of many contemporary effective diagnostic and therapeutic measures, this principle must be understood as a challenge for a benefit-risk evaluation of each medical intervention. This evaluation, related to the individual patient and put into effect by him- or herself as well, will increasingly be influenced by his or her social and especially economic context.

Example
Many physicians see the efficiency of current antidementive drugs as too low for their application, and accordingly such drugs will not be paid for, e.g., by the National Health Service in the United Kingdom.

Such application of limited resources for the best possible, i.e., evidence-based, help for the largest possible number of ill people will be justified by the principle of **justice**.

The principle of medical **confidentiality** is also already stated in the Hippocratic Oath. Because without seriously taken confidentiality, medical action is endangered in its core, since without it the necessary openness for the appropriate recognition of the disease and its determining conditions as well as an adequate understanding of the patient is almost impossible. This is particularly valid in a time in which the psychiatrist faces manifold demands for disclosure of data about his patient, by his team and other physicians, relatives, and students, as well as by health insurance companies or even agencies of research control. Confidentiality is especially important for people with mental illness in view of the widely existing and

stigmatizing skepticism they have to face and also due to the often specific vulnerability of their interpersonal relations.

Example
The candidate for the American vice presidency in 1972 lost his candidacy because the public was informed that he had been treated transiently in a psychiatric outpatient facility.

Ethical Rules

To sensitize psychiatrists for such ethical implications of their actions in order to protect the patients, a set of ethical rules, codices, etc. have been developed, both generally and specifically, for psychiatric practice as well as for psychiatric research, e.g., the following:

- Declaration of Madrid (1996) of the World Psychiatric Association (WPA) (1996) with its revisions, the latest one currently being formulated
- Opinions of the Ethics Committee on the Principles of Medical Ethics – With Annotations Especially Applicable to Psychiatry (2009) of the American Psychiatric Association (APA) (2009)
- Good Psychiatric Practice (2009) of the Royal College of Psychiatrists (RCP) (2009)

Specific guidelines for research with mentally ill people are, e.g., the following:

- Ethics of psychiatric research (2011), a position statement of the British Royal College of Psychiatrists (2011), related to:
- Guidelines for Researchers and for Research Ethics Committees on Psychiatric Research Involving Human Participants (2000) of the Royal College of Psychiatrists (2000)

Ethical guidelines especially for medical research are:

- The Declaration of Helsinki (1964) with its revisions up to 2008 of the World Medical Association (WMA) (2008).
- Chapter V, Articles 15–18 of the Convention on Human Rights and Biomedicine (1997) of the European Council (Europarat 1997).

  Psychiatrically relevant is §17.2 insofar as it declares research with ill persons who are not capable of giving informed consent as acceptable in specific exceptional cases. This has led to a highly controversial and emotional public debate in Germany (de Wachter 1997; Helmchen 1998, 2008, 2010; Maio 2010).
- European Textbook on Ethics in Research (2010) of the European Commission's Directorate-General for Research (European Commission's Directorate-General for Research 2010).

It should be noted that only laws and related governmental regulations are legally binding; declarations, recommendations, and other statements are not. Nevertheless the ethical norms of the latter, particularly those of the Declaration of Helsinki or of

national licensing authorities, have shaped the performance of clinical researchers as well as the acceptability of human research and its limitations and have influenced legislature.

Examples

The following contributions are only a very few examples of fairly different situations of psychiatric action that show the ethical complexity or even conflict-laden problems with which the psychiatrist is confronted. The examples are centered on specific aspects of respect for the autonomy of the patient and on the assessment of the risk-benefit relationship of each intervention – at present intensely discussed ethical questions.

The capacity of self-determination is psychiatrically relevant since mental illness may impair not only autonomy directly but also the social relationships of the person with a mental disease. This has now become much more relevant after the passage of the UN Convention on the Rights of Persons with Disabilities (CRPD) (2006), because the relationship of the capacity to consent to human rights is under discussion.

One aspect of self-determination is the informed consent as a binding requirement of each medical intervention. Not only sufficient information but also the patient's capacity to consent is required for its validity. Its assessment is important in order not to overload an incompetent patient with a responsibility that he cannot bear. However, its assessment is difficult and needs experience and accuracy, because the capacity to consent does not usually suddenly or completely disappear. Furthermore, it must be assessed with regard to a specific fact, and it depends on its complexity and meaning. Thus capacity to consent is to be assessed only gradually (in levels of strength) and in relation to the situation, but not generally as existent or not existent.

Szmukler and Ross describe the new development driven by service users towards respecting and enhancing self-determination of persons with mental illness both on the individual and on the institutional level. They recommend the use of advance statements, i.e., statements of what a service user wishes to be done and what not in case of a future episode of mental illness. They mention the controversies about the usefulness of such advance statements and discuss critically their limitations, e.g., the considerable uncertainty about circumstances under which such statements may be overridden – on which the contribution of Konrad and Müller goes into detail. Furthermore they point to the very new development of the participation of patients in some aspects of research such as setting the research agenda. Finally they argue for the abolition of a separate (risk-oriented) mental health legislation[2] in favor of a "fusion" legislation, based on impaired

[2]See also Callard et al. (2012).

decision-making capacity and best interests of the patient, irrespective of whether he or she has a mental or a somatic illness.

In contrast to this sign of a paradigmatical change towards the primacy of the patient's decision, Konrad and Müller discuss the existing legal situation and the uncertainties after the latest decision of the German Constitutional Court, which says that the existing law is insufficient with regard to clear criteria for compulsory treatment in case of compulsory admission. They also point to conflicts between ethical principles such as respecting the autonomy of the patient in case of refusal of a lifesaving treatment and the obligation of the physician as a guarantor; a series of rules has been established only if a person with full decisional capacity begins a hunger strike.

Another pertinent topic is addressed by the evaluation of the risk-benefit relationship of psychiatric interventions, i.e., the dimension of probability and thereby that of dealing with uncertainty and with aporias such as the impossibility to compare risks and benefits on the individual level with those on the societal level (Helmchen, chapter in this section). This is mainly valid for research projects. However, societal demands increasingly also influence the individual risk-benefit evaluation, mainly by economical and priority considerations.

These three examples may give only a glimpse of the breadth and complexity of ethical implications of psychiatry. Thus, the interested reader may find more examples in two books: Green and Bloch (2006) and Helmchen and Sartorius (2010).

References

American Psychiatric Association. (2009). *Opinions of the Ethics Committee on the principles of medical ethics – With annotations especially applicable to psychiatry*. Arlington: American Psychiatric Association.

Beauchamp, T., & McCullough, L. (1994). *Medical ethics. The moral responsibility of physicians*. Englewood Cliff: Prentice-Hall.

Callard, F. N. S. N., Arboleda-Flórez, J., Bartlett, P., Helmchen, H., Stuart, H., Taborda, J., & Thornicroft, G. (2012). *Mental illness, discrimination and the law: Fighting for social justice*. Chichester: Wiley-Blackwell.

de Wachter, M. A. M. (1997). The European convention on bioethics. *The Hastings Center Report, 27*(1), 13–23.

Europarat. (1997). Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine (No. 164) (CHRB-97).

European Commission's Directorate-General for Research. (2010). European textbook on ethics in research. http://wwweurosfaireprdfr/7pc/doc/1292233423_textbook_on_ethics_report_enpdf

Green, S., & Bloch, S. (2006). *An anthology of psychiatric ethics*. New York: Oxford University Press.

Helmchen, H. (1998). Research with incompetent patients. A current problem in light of German history. *European Psychiatry, 13*(Suppl. 3), 93s–100s.

Helmchen, H. (2008). Clinical research in the mentally ill. Ethical consideration. In F. Thiele, J. M. Fegert, G. Stock (Eds.), *Clinical research in minors and the mentally ill*. Europäische Akademie zur Erforschung von Folgen wissenschaftlich-technischer Entwicklungen. Bad Neuenahr-Ahrweiler 2008: 7–31.

Helmchen, H. (2010). Ethical guidelines in psychiatric research. *European Archives of Psychiatry and Clinical Neuroscience, 260*, 142–146.

Helmchen, H., & Sartorius, N. (2010). *Ethics in psychiatry*. Dordrecht/Heidelberg/London/New York: Springer.

Koch, H. G., Reiter-Theil, S., & Helmchen, H. (1996). *Informed consent in psychiatry*. Baden-Baden: Nomos.

Maio, G. (2010). Ethics of research with decisionally impaired patients. In H. Helmchen & N. Sartorius (Eds.), *Ethics in psychiatry* (pp. 421–436). Dordrecht/Heidelberg/London/New York: Springer.

Royal College of Psychiatrists. (2000). Guidelines for researchers and for research ethics committees on psychiatric research involving human participants. In: Royal College of Psychiatrists (Ed.), Council report CR82, London.

Royal College of Psychiatrists. (2009). Good psychiatric practice. In: Royal College of Psychiatrists (Ed.), College report CR154, London.

Royal College of Psychiatrists. (2011). Ethics of psychiatric research. In: Royal College of Psychiatrists (Ed.), Position statement PS02/2011, London.

UN. (2006). Convention on the Rights of Persons with Disabilities (2006) (CRPD-06) http://www2ohchrorg/english/law/disabilities-conventionhtm

World Medical Association. (2008). Declaration of Helsinki (1964/2008). http://wwwwmanet/en/30publications/10policies/b3/17cpdf

World Psychiatric Association. (1996). Declaration of Madrid on ethical standards for psychiatric practice. http://wwwwpanetorg/detailphp?section_id=5&content_id=48

Strengthening Self-Determination of Persons with Mental Illness

56

George Szmukler and Diana Rose

Contents

Abstract

"Self-determination" is taken to mean a freedom from forms of control or coercion deriving from external limitations imposed through common treatment practices and social institutions. The relationship between self-determination and discrimination is noted.

G. Szmukler (✉)
King's College London, Institute of Psychiatry, London, UK
e-mail: george.szmukler@kcl.ac.uk

D. Rose
Health Service and Population Research Department, King's College London Institute of Psychiatry, London, UK
e-mail: diana.rose@kcl.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_55,

Four domains of mental health practice are considered in which patient self-determination can be enhanced. First, treatment at the individual level, how, in everyday practice, patient self-determination may be respected and enhanced in the relationships between the patient and members of the clinical team. The "recovery movement" and "advance statements" are examples. Second, how service user involvement may shape the services offered to patients at a local institutional level, a national level, and even an international level. Third is a recent development, the involvement of patients as collaborators in the conduct of research, as distinct from participating as research "subjects." Here there is the opportunity to influence the research agenda, including the questions that should be asked and thus what knowledge is deemed important in mental health care. Fourth, constraints that limit patient self-determination are examined. These include the range of "treatment pressures" exercised on patients reluctant to accept a proffered treatment, including those that can be termed "coercive." Powerful sociopolitical pressures, generated to a large degree by stereotypes of mental illness, act against patient self-determination. The way in which mental health law is constructed reflects these sociopolitical influences. Conventional mental health legislation discriminates against persons with "mental illness" and fails to respect their autonomy to the same degree as persons with "physical illness." Proposals for mental health law reform are discussed.

Introduction

A pragmatic approach will be taken to the meaning of *self-determination* in this chapter. The Oxford English Dictionary defines the term as "determination of one's mind or will by itself towards an object." There is clearly a relationship to the idea of *autonomy*. Without looking in detail at the meaning of this contentious concept, a distinction drawn by Bolton and Banner (2012) between two general spheres of application is helpful for our purposes. One "highlights the conditions under which a person's desires, beliefs, reasons and actions can be considered as originating in or belonging to the self, as authentic in this sense – as opposed to having some other origin." In contrast to this "personal" meaning is one that is "social" or "political," "in which context it means freedom of the individual citizen to carry on with his or her own affairs, in particular freedom of the individual from state control."

This distinction applies equally to *self-determination*. Common to both meanings of "self-determination" is a freedom from some form of control or coercion. In the "personal" sense, that control results in some way from limitations imposed by, for example, a mental illness; in the "social-political" sense it derives from external limitations imposed through the operations of social institutions. The latter will be the main focus. However, it is very likely that enhancing self-determination in the sociopolitical sphere (e.g., by listening to and respecting the person or by not excluding him or her from community participation, resources, and power) reduces the undermining of self-determination by the mental illness in the personal sphere.

A connection between self-determination in the sociopolitical sphere and "discrimination" should also be noted. To deny civil-political (sometimes called

"negative" rights) or socioeconomic ("positive" rights) to persons with mental illness on grounds that are clearly unfair is both discriminatory and an unwarranted interference with self-determination.

Outline of the Chapter

First we shall consider how patient or service user[1] self-determination may be enhanced rather than restricted in the way mental health care is practiced and its services organized. Treatment at the individual level shall be examined – how, in everyday practice, patient self-determination may be respected and enhanced in the relationships between the patient and members of the clinical team. Next, we shall examine how service user involvement may shape the services that are offered to patients at a local institutional level, up to a national level and even an international level.

Then a recent development that may come to shape mental health care at a fundamental level will be briefly described – the involvement of patients as collaborators in the conduct of research, as distinct from participating as research "subjects." Here there is the opportunity to influence the research agenda, including the questions that should be asked and thus what knowledge is deemed most important in mental health care.

Finally, constraints that limit patient self-determination shall be examined. These include the range of "treatment pressures" that may be exercised on patients who are reluctant to accept a proffered treatment, including measures that can be termed "coercive." Powerful sociopolitical pressures, generated to a large degree by stereotypes of mental illness, act against patient self-determination. The way in which mental health law is constructed reflects these sociopolitical influences. We shall examine how conventional mental health legislation discriminates against persons with "mental illness" and fails to respect their autonomy to the same degree as persons with "physical illness." Proposals for mental health law reform aimed at dealing with such discrimination will be discussed.

Thus self-determination will be discussed under four headings:

1. Treatment at the individual level
2. Treatment at the level of service design or function
3. Research in mental health: involvement in research; setting the research agenda
4. Treatment pressures, coercion, and the law relating to mental health

[1]A point of terminology should be noted here. There are various terms that have been used to describe those who use mental health services, for example, "patients," "consumers," "clients," "survivors," and "service users." These terms may differ in the sets of values they imply. We shall use the terms "service user" and "patient" interchangeably, as they may be, at the present time at least, the least controversial and ones with which the coauthors of this chapter are reasonably content.

Treatment at the Individual Level

The treatment wishes and preferences of patients with mental illness, especially those treated in public mental health services, have not achieved a level of regard comparable to those of patients in other medical specialties. It is worth noting, for instance, that the history of *voluntary* inpatient treatment in a public mental hospital is relatively recent. Jurisdictions began to introduce "voluntary" status in the early twentieth century. For example, in the Netherlands this became possible in 1904; in Victoria, Australia, in 1916; in England not until the passing of the Medical Treatment Act in 1930 (although there were ways of achieving it earlier); and in Italy, not until the passing a special decree in 1968.

Some approaches, mainly recent, to enhancing patient self-determination will now be considered.

The "Recovery Movement"

The past two decades have seen a significant change in accounts of the philosophy of care, most evident in the Anglophone countries. Whether a change will be largely at the level of rhetoric or whether it signals a real change in how the patient's voice is heard remains to be seen. The change in approach is known as the *recovery movement*. The precise meaning of "recovery" in this context varies from one authority in the field to another; but central to the notion is an enhancement of patient self-determination and the view that the goals of treatment are to be largely set by the patient, not the clinician (Slade 2009). Conventionally in medicine, "recovery" means restitution to the pre-illness state, with the loss of symptoms and dysfunctions. "Recovery" in the "recovery movement" sense has a different meaning. An influential paper by Anthony (1993) gives an often-quoted definition:

> a person with mental illness can recover even though the illness is not 'cured'. Recovery is a way of living a satisfying, hopeful, and contributing life even with the limitations caused by illness. Recovery involves the development of new meaning and purpose in one's life as one grows beyond the catastrophic effects of mental illness.

"Recovery" (or "personal recovery") in this sense comprises both a "journey" (or "being in recovery") as well as an "outcome," but an outcome that does not necessarily require symptom loss. A key goal is to find a new way of living, through self-agency. It involves regaining a sense of self, becoming empowered to take control of one's life, and combining optimism for a better future with an acceptance of the past.

This change in philosophy requires a change in some of the traditional values underlying mental health care. Slade and Davidson (2011) describe it thus:

> Previous preoccupations (e.g. risk, symptoms, hospitalisations) become seen as a subset or special case of the new paradigm. By contrast, what was previously of peripheral interest (i.e. the patient's perspective) becomes central. This involves a reversal of some traditional clinical assumptions. Mental illness is a part of the person, rather than the person being

> a mental patient. Having valued social roles improves symptoms and reduces hospitalisation, rather than treatment being needed before the person is ready to take on responsibilities and life roles. The recovery goals come from the patient and the support to meet these goals comes from the clinician among others, rather than treatment goals being developed which require compliance from the patient. Assessment focuses more on the strengths, preferences and skills of the person than on what they cannot do.

Slade and Davidson suggest there are four ways in which clinicians can support an individual's recovery: fostering relationships, promoting well-being, offering evidence-based treatments to achieve the patient's recovery goals, and tackling social exclusion and stigma. They argue that mental health services need to be oriented around these recovery support tasks if they are to support recovery. Other writers have emphasized the importance of a reawakening of hope, of peer support and narratives of recovery, of a "social model" of disability, and of the need for risk taking (e.g., Repper and Perkins 2003; Kelly and Gamble 2005).

Although the "recovery" concept has now found a place in government mental health policy in the USA, UK, Australia, and New Zealand, doubts have been expressed concerning its full implementation. Changes in practice are always difficult to achieve, and those involving substantial shifts in underlying values, especially so. Concepts like "recovery" that have their origins in the service user movement can lose their vitality when "standardized" for institutional settings. Bonney and Stickley (2008) following a detailed review of the literature see some major limitations:

> With such a current focus upon risk in Western society, it is virtually inconceivable that statutory health care providers will ever fully embrace the recovery paradigm that involves self-management and has choice, hope, freedom and autonomy at its core. Furthermore, these values are extremely difficult to measure in a system that revolves around targets and outcomes. While there may be workers within the system who genuinely subscribe to recovery principles, they will struggle to practise according to those values in a system that pays only lip-service to a philosophy that is very dependent upon human values and beliefs.

Some patients might opt for seeking recovery entirely outside conventional mental health services. The availability of such facilities is highly problematic, because of funding as well as the constraints identified above. A possible unintended consequence of the "recovery" idea is that if "recovery" places a high value on self-management and if a good outcome for the patient can be achieved without cure, there may be a temptation, especially when finances are tight, to cut back on some services. Some patients have expressed concerns about a new set of normative demands that they fear they may be unable to meet and that failure to embark on what is seen as a "recovery journey" might leave them unsupported by both peers and services.

Despite these limitations there is an important value in the "recovery" approach. In some ways, the "recovery movement" positions patients with a mental illness in the same place as the generality of medical patients. It offers the same scope for choosing (or rejecting) treatment in accordance with their values and preferences. What appears to be a radical change in treatment philosophy is in many respects the removal of long-standing discrimination in mental health care if a patient's goals

should be at variance with those of the clinician. However, use of the word "recovery" may be less than optimal. It creates confusion in the minds of mental health professionals who invariably argue that recovery (albeit in the medical sense) has always been their aim; to suggest otherwise may appear to be an unfounded criticism and lead to resistance. Perhaps a term such as "reinstatement" or "re-engagement" might have been better.

Advance Statements

A specific clinical tool aimed at enhancing self-determination is the "advance statement" (AS). An AS allows a patient, when well, to state treatment preferences in anticipation of a time in the future when, as a result of the effects of illness, he or she may not be capable of making or expressing treatment decisions. The anticipated loss of decision-making capacity usually occurs during a relapse of a psychosis. An important aim of an AS is to give patients greater say over their treatment at a time of crisis and thus to reduce the need for "coercive" interventions.

A typology of ASs has been described by Henderson et al. (2008a). They vary along the following dimensions: whether patient or service provider led, whether legally binding or not, and whether facilitated by a person independent of the clinical team. At least three major types of AS have been described: "crisis cards" (CCs), "joint crisis plans" (JCPs), and "psychiatric advance directives" (PADs).

In a **crisis card** (CC) patients state their treatment wishes or nominate a person familiar with their preferences, without involvement of their treatment team. It has its origins in the service user self-advocacy movement and is entirely patient led. In some jurisdictions, for example, under the Mental Capacity Act 2005 (MCA) in England and Wales, stated treatment refusals now have legal force but can be overridden by mental health legislation. CC uptake has been very limited.

In contrast to CCs, the **joint crisis plan** (JCP) is the product of a semi-structured discussion between patient (supported by a relative, friend, or advocate) and the clinical team. The JCP reflects an agreement on what measures should be taken if a relapse should occur. A facilitator, independent of the clinical team, ensures that the patient's voice is given prominence (Sutherby et al. 1999; Henderson et al. 2004). A JCP is not legally binding (although as noted above specific treatment refusals may have legal force in some jurisdictions). However, the clinical team makes it explicit that while it will attempt to honor the terms of the JCP, compliance with all aspects cannot be guaranteed. While the JCP is provider initiated, it achieves an agreement in which the patient's voice is key, and is formulated in the patient's words. The JCP's specificity of content, based on a detailed reconstruction of past illness episodes and the treatment offered, is an important advantage since relapse tends to take a more or less stereotyped form.

A randomized controlled trial of JCPs has been conducted in South East England (Henderson et al. 2004). Almost 40 % of patients who were eligible took up the

opportunity to complete a JCP. Eligible patients were those with a diagnosis of a psychotic illness or bipolar disorder and who had had at least one admission in the previous 2 years. Compulsory admissions over a 15-month period were halved compared to a control group. On follow-up patients said their agreements were given freely and without pressure and that they felt more in control of their mental health problems as a result of making a JCP (Henderson et al. 2009). We await the results of a large replication study that is soon to be completed.

Psychiatric advance directives (PAD) may take three forms: (i) specified treatments that are refused or requested; (ii) statements about values, attitudes, or preferences to guide those making treatment decisions for the patient; and (iii) nomination of a person to act a “substitute” (or “proxy”) decision-maker. A PAD assumes the patient had decision-making capacity when it was made and that the circumstances in which the PAD is triggered are those that were anticipated.

PAD legislation now enacted in a number of states in the USA ostensibly makes such directives legally binding. Nevertheless, there is considerable uncertainty about circumstances when PADs may be overridden (Swanson et al. 2006a). Civil commitment legislation may do so if it specifically authorizes involuntary treatment. However, in those states where a separate step requiring the patient’s consent before treatment can be given following detention in hospital (the detention being based, e.g., on a separate dangerousness standard), a PAD may prevail. Some PAD statutes, for example, in Pennsylvania, stipulate that the patient’s wishes may be overridden if they violate “generally accepted community practice standards.” How this should be interpreted is not clear. It is noteworthy that the evidence on PADs to date shows that refusal of all treatment alternatives is rare (Swanson et al. 2006b). A controversial issue is whether a PAD making a treatment refusal very likely to lead to a life-threatening risk should be respected or not. This is too big a subject to be considered here.

A **facilitated PAD** (F-PAD) was introduced following surveys showing that despite an apparently widespread appeal of PADs, few patients actually make one, probably because of their complexity. In an F-PAD a trained facilitator explains what a PAD involves and, if the patient chooses to opt for one, assists with its completion (Swanson et al. 2006b). Sometimes the service provider is brought into the process. A randomized controlled study of F-PADs found that facilitation resulted in a highly significant increase in the number of patients who decided to make a PAD (61 % vs. 3 % of controls). At 1-month follow-up, those with an F-PAD reported a better therapeutic alliance with their clinicians and were more likely to say that they received the services they needed (Swanson et al. 2006b). A later report from this study found that the number of “coercive interventions” (e.g., police involvement, involuntary treatment, seclusion) over the succeeding 2 years for those who made a PAD was considerably fewer compared to patients who chose not to make a PAD (Swanson et al. 2008).

Thus there is evidence that some forms of AS can reduce coercive interventions and improve outcomes from the patient’s perspective. Patients sense of self-determination may be enhanced; from a practical point of view, the information

contained in an AS may ensure that the most appropriate treatment is given when information from other sources is unavailable. JCPs suggest that patients can effectively voice their treatment wishes outside a legal framework.

Hospitalization

Admission to a psychiatric hospital is usually an especially distressing experience, exacerbated by coercive treatment interventions. However, there is evidence from a number of non-controlled studies that the use of seclusion or of physical restraint can be significantly reduced (Gaskin et al. 2007); Hallerstein et al. (2007). Approaches have been varied, are multiple, and have usually involved system changes – for example, state policy and regulation changes aimed at reducing seclusion, monitoring and analyzing episodes, strengthening leadership; staff education, changing the ward environment, increasing staff to patient ratios, creating special emergency response teams, and treating patients as active participants in interventions to reduce seclusion. A study attempting to increase the involvement of inpatients in planning their treatment did not affect the "perceived coercion" experienced by the patients (Sorgaard 2004). More research in this area is clearly warranted.

Treatment at the Level of Service Design or Function

Here developments in the UK will serve as an example, but it should be noted that the extent of service user involvement varies hugely, even within and between countries with well-developed mental health policies and services. New legislation in 1990 (Community Care Act) (Department of Health and Social Security 1990) ushered in the policy of recipients of health and social care services being consulted about the way in which these services should be configured and delivered locally. Since then, and incrementally, successive governments have pursued service user involvement in service design and function (Department of Health 2000, 2011). The most recent is the UK government's 2011 policy statement "No Health Without Mental Health" which reiterates that patient choice should be at the heart of care and promises "no decision about me without me." It should perhaps be noted that this is a reprise on a long-standing slogan of the disability movement "nothing about me without me."

Patients often serve on committees charged with developing or reconfiguring services and may also sit on appointment panels for some categories of staff. A major mechanism through which the government now seeks to increase choice over care is through the introduction of personal budgets (PBs). Instead of funding services through blocks of money to health-care providers, service users are to be given a PB to buy care that suits their needs. This, it is said, will ameliorate ghettoization and dependency (features of traditional day services for example) and enhance service user autonomy.

A further mechanism for service users to influence local policy is to allocate reserved places on the governing boards of hospital and community health-care organizations. In this way, service users have a voice in local decision-making and contribute to the policies and strategy of the facility from which they receive services. Some user "governors" conceive of themselves as "representing" their constituency, but others act as individuals with no mandate.

A question arises as to whether there has, in the decades since 1990, been a significant transition in the ways in which service users can affect service configuration and delivery. The traditional method was through *collective* organization where groups of service users, often user led, consulted or campaigned to bring about change. They were often criticized by service providers as "unrepresentative" (Rose et al. 2010). With the shift to PBs and the development of user governor structures, it may be argued we are seeing something much more individualized. This befits an ideology of the moment, but it may have shortcomings in practice. In a climate of financial austerity, for example, only a select few have access to PBs and many have no services at all. The arguments about the parlous nature of ghettoization and dependency may mean, in effect, that many rely on inappropriate or underfunded provision. The ethical requirement for self-determination in such circumstances may ring hollow and amount to little more than an obligation to be a free and liberal subject.

Partners in Research

"Service user (or patient) involvement in research" refers to the active involvement of service users as partners or collaborators in the conduct of research – helping to design, deliver, and disseminate research. It is distinct from service user participation as research "subjects." Service user involvement in research is a recent development, starting in the mid-1990s. In the UK it is now supported by a number of Department of Health policies and is becoming increasingly developed through its Clinical Research Networks, set up to provide logistic support for high-quality medical research (http://www.nihr.ac.uk/infrastructure/Pages/infrastructure_clinical_research_networks.aspx).

Demonstrating evidence for an impact of service users on research is complex and still in its infancy. Involvement may occur at different stages of the research enterprise and with different expectations – for example, improving the practical aspects of the study to enhance recruitment (Donovan et al. 2002), choosing meaningful outcome measures (Crawford et al. 2011), and improving the quality of data acquisition or interpretation (Gillard et al. 2010; Rose et al. 2011). These require different methods of evaluation. While evaluation has not generally reached this level of sophistication, evidence of impact is accumulating. For example, Vale et al. (2012), surveying service user involvement in studies run by the Medical Research Council Trials Unit, found that researchers reported a range of positive impacts on the research and researchers, including

improved credibility, design and quality, trial recruitment, and dissemination. Other evidence, mainly from case studies, is consistent with this report (Staley 2009).

But besides the issue of quality of research, there is also a strong moral case. The framework is well articulated in Kitcher's (2001) notion of a *well-ordered science*. It revolves around how a research agenda should be set? While traditionally this has been decided by "elites" – communities of scientists, perhaps in association with a privileged group of outsiders, such as the funders of research or industry – Kitcher argues this is not an acceptable basis for setting a research agenda. Science is clearly embedded in society, and since it potentially affects the lives of all, a range of voices needs to be heard. However, what he terms "vulgar democracy," baldly following the wishes of the general public, is not the answer; it leads to an emphasis on current "hot topics," usually with a short-term focus. The basis for a *well-ordered science* favored by Kitcher is "enlightened democracy," founded on "tutored preferences." Representative citizens in dialogue with scientists would examine a particular problem, what is known about it, how it might be researched, and the problems and prospects for knowing more. They will thus understand the goals, methods, and limitations of a particular research approach for solving a particular problem. They become deliberators with "tutored preferences" who can engage in discussions with the scientists, funders, government, and each other about research priorities for the common good.

This "tutoring" process is dialogical; the scientists will learn from the relevant nonscientists about the personal meanings and sociopolitical implications of their knowledge. Ideally then, the research agenda will be set by informed deliberators who will determine the problems to be researched, the kinds of science best suited to solve them, and the constraints (e.g., ethical, financial) that need to be imposed. In medical research, service users are a key stakeholder group.

A well-ordered science along these lines is obviously an ideal. However, service user involvement in the conduct of research has generated practices that can move us in that direction. Certainly, there are obstacles in understanding each other's viewpoints. For lay persons the science can be difficult to understand, as can be the complex workings of academia, while for the researcher, the experience of mental illness can be difficult to appreciate. Progress towards a "well-ordered science" in mental health may be especially challenging; for instance, so much is contested, even the term "illness" itself and thus of the value of the application of a range of scientific approaches to mental health "problems."

Nonetheless, through an organization such as the UK Mental Health Research Network, one has seen the research agenda being increasingly influenced by patients (Staley 2012; Staley et al. 2012). All studies adopted for support have their involvement, and it is apparent that service users are increasingly engaged in decisions concerning the aims, methods, and design of studies. Serving as members of project management or steering committees is now common. We anticipate that over time, service users will have an increasing influence in setting the research agenda and perhaps in creating new kinds of knowledge.

Coercion and the Law Relating to Mental Health Care

In this section constraints on patient self-determination will be examined that arise from the actions of mental health professionals, the construction of the mental health service "system," and measures sanctioned by mental health law.

Coercion

One can think in terms of a hierarchy of *treatment pressures* placed on reluctant patients: "persuasion," "interpersonal leverage," "inducements" (or offers), threats, compulsion (Szmukler and Appelbaum 2008). Persuasion, an appeal to reason, and "interpersonal leverage," appearing to be disappointed by the patient's refusal in the context of a valued relationship, are not especially problematic, though one should be aware of the subtle influence of the latter. Needing closer attention are the last three. It can be argued that *coercion* in a technical sense enters at the point of "threats." In contrast to an inducement, a threat is a conditional proposition that if not accepted by the patient makes the person, in an important sense, usually according to an accepted set of values, worse off. "If you do not take your medication, then you will be admitted to hospital involuntarily" is a threat. Rejecting the proposition leads to one being worse off than one was in the "baseline" position, prior to the proposition having been made – being deprived of one's liberty. An offer, on the other hand, if not accepted does not make the person worse off. "If you take your medication, you can join a patient group on a day trip to a lovely seaside town"; rejection of the proposition leaves the person's baseline position unchanged; it is the same as if the proposition were never made.

It is worth noting that leverage (threats) of various kinds, short of compulsion, are common in psychiatric practice – a study of community patients in Oxford, for example, found that up to 30 % of patients had experienced these in the past (Burns et al. 2011).

While inducements are not on this account coercive, they may still be problematic. An example is paying patients with a psychosis to take medication (Priebe et al. 2010). Most clinicians have a moral intuition that this is in some way problematic. Factors that may contribute to this feeling are exploitation (the clinician in some way takes unfair advantage of the patient), unfairness (why should the unwilling patient be paid, but not the prudent one), and an exchange that involves very different domains of value that cannot be measured on the same metric (e.g., selling a child). The higher value – respect for the person, something that mental health services should be enhancing not diminishing – is somehow degraded (Szmukler 2009).

Moving up the hierarchy to "compulsion," there is a striking observation to be made. This is the extraordinary variation in its use between countries – and within countries. There was thirtyfold variation in involuntary admissions to hospital

across a group of European countries in 1998–2000 (Salize and Dressing 2004). Within countries, large variations have also been noted, for example, fivefold in Norway, despite the same legislation and same service structures. Just as significant are changes in rates of involuntary hospitalization over time. In England and Wales, compulsory admissions to hospital increased by 63 % between 1984 and 1996 (Hotopf et al. 2000). Numbers in Sweden declined over the same period (Kjellin et al. 2008). Large variations in rates of application of coercive measures following admission are also evident, for example, in the use of seclusion and physical restraint. In a 10-country study, this varied between 21 % and 59 % in first 4 weeks of involuntary admissions. Rates also vary for different ethnic groups in a country.

Politics, culture, and local "custom and convention" are highly influential. A positive inference that might be drawn from an otherwise troubling set of findings is that in many places there may exist scope for significant reductions in countries with high rates.

Recent studies have also shown that patients subject to involuntary admission often fail, in retrospect, to endorse the action. A recent study in England found this to be the view of 60 % of patients followed up 1 year later (Priebe et al. 2009).

While coercion and some types of offer may constitute, on the face of it, a wrong, it does not mean that it cannot be justified in some circumstances – appropriate criteria for involuntary treatment, for example, define those. One approach, for example, is based on a "capacity-best interests" test. We return to the subject below.

Discrimination and Mental Health Law

There is a strong argument (Dawson and Szmukler 2006; Szmukler et al. 2010) that mental health legislation of the conventional kind discriminates unfairly against persons with mental illness and undermines rights to self-determination (or "autonomy") that, by contrast, are un-controversially possessed by patients with "physical disorders." Unlike the situation in all other medical specialties, such mental health law allows persons with mental illness to be treated against their will even if they retain decision-making capacity (or capability) – that is, they are able to understand and retain the information relevant to their illness and the treatment (e.g., the likely consequences of having or not having the treatment), to appreciate its relevance to their predicament, and to be able to reason with it in the light of their values and life choices, thus deciding whether to undergo the treatment or not.

The necessary criteria in most jurisdictions for treating a patient without their consent in general medicine is that the patient lacks such capacity and that the treatment is in some sense in the person's "best interests." The latter criterion is variously understood, but the strongest case can be made for "best interests" referring to what the patient would have chosen in this predicament if they had capacity – based perhaps on an advance statement (such as a JCP) or on

a consideration by those who know the patient well of his or her values and life choices that would be relevant. The usual criteria for persons with mental illness, on the other hand, are based on (i) the presence of a "mental disorder" and (ii) the presence of a "risk" of some kind to the person's health or safety or to the safety of others. These risk-based justifications, it is suggested, reflect deeply entrenched stereotypes of people with a mental illness – that they are intrinsically less than competent persons because of their illness or that they are dangerous.

In order to deal with this discrimination, Dawson and Szmukler (2006) have proposed a major reform of existing mental health legislation. Indeed, they argue for the abolition of separate mental health legislation altogether, in favor of a single, comprehensive statute with a single set of criteria governing involuntary treatment of all persons, no matter what the health setting – general medicine, surgery, or mental health. The proposal is based squarely on an impairment of decision-making capability as the essential gateway to a consideration of involuntary treatment, followed by a "best interests" assessment. The same legislation could also readily cover social care where an action such as a nursing home placement is being considered that is contrary to the person's wishes. The law that is being proposed is termed a *fusion law* as it combines, on the one hand, the respect for patient autonomy that one finds in general medical law or "capacity" legislation based on impaired decision-making capability and, on the other hand, the regulation of detention and forced treatment, once instigated, that is well covered by civil commitment law, for example, by specifying how it is to be authorized, by whom, for how long, where, how appeals are to be made, and how often. Capacity legislation (as in the Mental Capacity Act 2005 in England) usually fails to do this adequately.

Involuntary treatment in the community has been a highly controversial recent development in mental health care. The *fusion law* approach would be compatible with involuntary community treatment provided the patient lacked capacity and the treatment was in their best interests. Another important difference compared to current risk-based legislation is that it would necessarily cease once capacity was recovered.

A capacity-best interests approach may also provide a framework for considering the justification for "coercive" interventions short of involuntary treatment (Szmukler and Appelbaum 2008).

UN Convention on the Rights of Persons with Disabilities (CRPD)

This convention, signed by 153 states and ratified by 119 as of August 2012 (United Nations 2006), may have a far-reaching effect on involuntary treatment. Though *disability* is not defined, at least some forms of mental illness, especially those that are serious and long term, are taken by the Committee on the Rights of Persons with Disabilities as comprising a disability. The CRPD adopts a social model of disability, where a disability has as much to do with the social responses and accommodations to a person with an impairment, as the impairment itself.

Notably, and controversially, the CRPD omits any mention of "substitute decision-making." The model is one of *supported decision-making* – the assumption is that people with disabilities can articulate an authentic preference if adequately supported to do so. Article 12 states that people with disabilities have "legal capacity on an equal basis with others." Not only do they hold rights, but they, like everyone else, have the right to exercise or act on those rights. The right to liberty (Article 14) states: "the existence of a disability shall in no case justify a deprivation of liberty." The UN High Commissioner for Human Rights (2009) has interpreted the intention as follows: "[48.]. . . unlawful detention encompasses situations where the deprivation of liberty is grounded in the combination between a mental or intellectual disability and other elements such as dangerousness, or care and treatment. Since such measures are partly justified by the person's disability, they are to be considered discriminatory and in violation of the prohibition of deprivation of liberty on the grounds of disability. . ." That is, if the disability comprises an arm of a set of criteria for involuntary treatment, it is discriminatory. "The legal grounds upon which restriction of liberty is determined must be de-linked from the disability and neutrally defined so as to apply to all persons on an equal basis." On this interpretation, States must repeal such laws; but such laws, as noted in the previous section, are almost the norm.

It has been argued that involuntary treatment and substitute decision-making are incompatible with the CRPD (Bartlett 2012; Minkowitz 2011). Some recent observations by the Committee on the Rights of Persons with Disabilities indicate that mental health law and discrimination will be the subject of special scrutiny. (See, e.g., Committee on the Rights of Persons with Disabilities. Sixth Session 19–23 Sept 2011. Concluding observations).

It can be argued that an involuntary treatment law that is decision-making capability and best interests-based, and applies to both physical and mental illness, such as the "fusion law," would be "disability neutral." Such law applies equally to the well person who, for example, suffers a head injury in an accident and to the person with schizophrenia who experiences a relapse of their illness. However, the "fusion law" approach may need to be reframed in order to emphasize supported decision-making and to recast the concepts of "decision-making capability" and "best interests" in terms of respect for, and facilitation of, the person's "will and preferences" (Szmukler et al. 2014). We wait to see how the CRPD will be interpreted, but it is likely to significantly change the discourse around involuntary treatment.

Although this section has focused on limitations on patient self-determination as a result of involuntary treatment, further limitations arise from failures to ensure equal socioeconomic rights for persons with mental illness. These are addressed in the CRPD and include, for example, the right to home and family life (Art. 23), the right to education (Art. 24), and rights to health (Art. 25) and habilitation and rehabilitation (Art. 2). Some of these rights have been reframed so as to have particular relevance to people with disabilities: the right to nondiscrimination (Art. 5), the right to independent living and community

inclusion (Art. 19), the right to personal mobility (Art. 20), the right to work and employment (Art. 27), the right to participation in cultural life (Art. 30), and the right to be free from exploitation and abuse (Art. 16).

What has been presented here is obviously a Western, high-income country, perspective on mental health legislation. However, it should be noted that around 25 % of the world's population live in states that have no mental health legislation and that in many states that do have such legislation, it is in any case not fully implemented.

Conclusions

Four domains have been considered where developments over the past decade or so seem to promise an enhancement of self-determination for persons with a mental illness – clinical practice, service design, the conduct of research, and mental health law. That these changes have occurred across a broad front suggests that there has been a change in attitudes and policies at a fairly deep level, and is encouraging. However, in each of the domains, our discussion has indicated that obstacles remain and that rhetoric may outstrip practice. It would be foolish to think that such long history of profound discrimination and stigmatization can be easily undone.

Cross-References

▶ Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
▶ Ethics in Psychiatry

References

Anthony, W. A. (1993). Recovery from mental illness: The guiding vision of the mental health service system in the 1990s. *Psychosocial Rehabilitation Journal, 16*, 11–23.

Bartlett, P. (2012). The United Nations Convention on the Rights of Persons with Disabilities and Mental Health Law. *The Modern Law Review, 75*, 752–778.

Bolton, D., & Banner, N. (2012). Does mental disorder involve loss of personal autonomy? In L. Radoilska (Ed.), *Autonomy and mental disorder* (pp. 77–99). Oxford: Oxford University Press.

Bonney, S., & Stickley, T. (2008). Recovery and mental health: A review of the British literature. *Journal of Psychiatric and Mental Health Nursing, 15*(2), 140–153.

Burns, T., Yeeles, K., Molodynski, A., Nightingale, H., Vazquez-Montes, M., Sheehan, K., & Linsell, L. (2011). Pressures to adhere to treatment ("leverage") in English mental healthcare. *The British Journal of Psychiatry, 199*(2), 145–150.

Crawford, M. J., Robotham, D., Thana, L., Patterson, S., Weaver, T., Barber, R., & Rose, D. (2011). Selecting outcome measures in mental health: The views of service users. *Journal of Mental Health, 20*(4), 336–346.

Dawson, J., & Szmukler, G. (2006). Fusion of mental health and incapacity legislation. *The British Journal of Psychiatry, 188*, 504–509.

Department of Health. (2000). *The NHS plan*. London: The Stationery Office.

Department of Health. (2011). *No health without mental health*. London: Department of Health.

Department of Health and Social Security. (1990). *National Health Service and Community Care Act*. London: The Stationery Office.
Donovan, J., Mills, N., Smith, M., Brindle, L., Jacoby, A., Peters, T., & Hamdy, F. (2002). Quality improvement report: Improving design and conduct of randomised trials by embedding them in qualitative research: ProtecT (prostate testing for cancer and treatment) study. Commentary: Presenting unbiased information to patients can be difficult. *British Medical Journal, 325*(7367), 766–770.
Gaskin, C. J., Elsom, S. J., & Happell, B. (2007). Interventions for reducing the use of seclusion in psychiatric facilities: Review of the literature. *The British Journal of Psychiatry, 191*, 298–303.
Gillard, S., Borschmann, R., Turner, K., Goodrich-Purnell, N., Lovell, K., & Chambers, M. (2010). "What difference does it make?" Finding evidence of the impact of mental health service user researchers on research into the experiences of detained psychiatric patients. *Health Expectations, 13*(2), 185–194.
Hellerstein DJ, Bennett Staub A, Lesquesne E (2007) Decreasing the use of restraint and seclusion among psychiatric inpatients. *Journal of Psychiatric Practice, 13*, 308–317.
Henderson, C., Flood, C., Leese, M., Thornicroft, G., Sutherby, K., & Szmukler, G. (2004). Effect of joint crisis plans on use of compulsory treatment in psychiatry: Single blind randomised controlled trial. *British Medical Journal, 329*(7458), 136.
Henderson, C., Flood, C., Leese, M., Thornicroft, G., Sutherby, K., & Szmukler, G. (2009) Views of service users and providers on joint crisis plans: Single blind randomized controlled trial. *Social Psychiatry and Psychiatric Epidemiology, 44*, 369–376.
Henderson, C., Swanson, J. W., Szmukler, G., Thornicroft, G., & Zinkler, M. (2008). A typology of advance statements in mental health care. *Psychiatric Services, 59*(1), 63–71.
Hotopf, M., Wall, S., Buchanan, A., Wessely, S., & Churchill, R. (2000). Changing patterns in the use of the Mental Health Act 1983 in England, 1984–1996. *The British Journal of Psychiatry, 176*, 479–484.
Kelly, M., & Gamble, C. (2005). Exploring the concept of recovery in schizophrenia. *Journal of Psychiatric and Mental Health Nursing, 12*(2), 245–251.
Kitcher, P. (2001). *Science, truth, and democracy*. New York: Oxford University Press.
Kjellin, L., Ostman, O., & Ostman, M. (2008). Compulsory psychiatric care in Sweden: Development 1979–2002 and area variation. *International Journal of Law and Psychiatry, 31*(1), 51–59.
Minkowitz, T. (2011). Prohibition of compulsory mental health treatment and detention under the CRPD. http://papers.ssrn.com. Accessed 17 April 2013.
Priebe, S., Katsakou, C., Amos, T., Leese, M., Morriss, R., Rose, D., & Yeeles, K. (2009). Patients' views and readmissions 1 year after involuntary hospitalisation. *The British Journal of Psychiatry, 194*(1), 49–54.
Priebe, S., Sinclair, J., Burton, A., Marougka, S., Larsen, J., Firn, M., & Ashcroft, R. (2010). Acceptability of offering financial incentives to achieve medication adherence in patients with severe mental illness: A focus group study. *Journal of Medical Ethics, 36*(8), 463–468.
Raboch, J., Kališová, L., Nawka, A., Kitzlerová, E., Onchev, G., Karastergiou, A., & Torres-Gonzales, F. (2010). Use of coercive measures during involuntary hospitalization: Findings from ten European countries. *Psychiatric Services, 61*(10), 1012–1017.
Repper, J., & Perkins, R. (2003). *Social inclusion and recovery*. London: Baillière Tindall. http://www.eu.elsevierhealth.com/Nursing/specialty/book/9780702026010/Social-Inclusion-and-Recovery/
Rose, D., Fleischmann, P., & Schofield, P. (2010). User perceptions of user involvement: A user-led study. *The International Journal of Social Psychiatry, 56*(4), 389–401.
Rose, D., Leese, M., Oliver, D., Sidhu, R., Bennewith, O., Priebe, S., & Wykes, T. (2011). A comparison of participant information elicited by service user and non-service user researchers. *Psychiatric Services, 62*(2), 210–213.
Salize, H. J., & Dressing, H. (2004). Epidemiology of involuntary placement of mentally ill people across the European Union. *The British Journal of Psychiatry, 184*(2), 163–168.

Slade, M. (2009). *Personal recovery and mental illness: A guide for mental health professionals*. Cambridge: Cambridge University Press.

Slade, M., & Davidson, L. (2011). Recovery as an integrative paradigm in mental health. In G. Thornicroft, G. Szmukler, K. T. Meuser, & R. E. Drake (Eds.), *Oxford textbook of community mental health* (pp. 26–33). Oxford: Oxford University Press.

Sorgaard, K. W. (2004). Patients' perception of coercion in acute psychiatric wards. An intervention study. *Nordic Journal of Psychiatry, 58*(4), 299–304.

Staley, K. (2009). *Exploring impact: Public involvement in NHS, public health and social care research*. London: National Institute for Health Research.

Staley, K. (2012). *An evaluation of service user involvement in studies adopted by the Mental Health Research Network*. London: UK Mental Health Research Network. http://www.mhrn.info/data/files/MHRN_PUBLICATIONS/REPORTS/Service_user_involvement_evaluation.pdf

Staley, K., Kabir, T., & Szmukler, G. (2012). Service users as collaborators in mental health research: Less stick, more carrot. *Psychological Medicine, 1–5*.

Sutherby, K., Szmukler, G. I., Halpern, A., Alexander, M., Thornicroft, G., Johnson, C., & Wright, S. (1999). A study of "crisis cards" in a community psychiatric service. *Acta Psychiatrica Scandinavia, 100*(1), 56–61.

Swanson, J. W., McCrary, S. V., Swartz, M. S., Elbogen, E. B., & Van Dorn, R. A. (2006a). Superseding psychiatric advance directives: Ethical and legal considerations. *The Journal of the American Academy of Psychiatry and the Law, 34*(3), 385–394.

Swanson, J. W., Swartz, M. S., Elbogen, E. B., Van Dorn, R. A., Ferron, J., Wagner, H. R., & Kim, M. (2006b). Facilitated psychiatric advance directives: A randomized trial of an intervention to foster advance treatment planning among persons with severe mental illness. *The American Journal of Psychiatry, 163*(11), 1943–1951.

Swanson, J. W., Swartz, M. S., Elbogen, E. B., Van Dorn, R. A., Wagner, H. R., Moser, L. A., & Gilbert, A. R. (2008). Psychiatric advance directives and reduction of coercive crisis interventions. *Journal of Mental Health, 17*, 255–267.

Szmukler, G. (2009). Financial incentives for patients in the treatment of psychosis. *Journal of Medical Ethics, 35*(4), 224–228.

Szmukler, G., & Appelbaum, P. (2008). Treatment pressures, leverage, coercion and compulsion in mental health care. *Journal of Mental Health, 17*, 233–244.

Szmukler, G., Daw, R., & Dawson, J. (2010). A model law fusing incapacity and mental health legislation. *Journal of Mental Health Law, Special Issue Ed, 20*, 1–140.

Szmukler, G., Daw, R., & Callard, F. (2014). Law on mental health consistent with the UN Convention on the Rights of Persons with Disabilities. *International Journal of Law and Psychiatry*.

UN High Commissioner for Human Rights. (2009). *Annual report to the General Assembly*. A/HRC/10/4, para 48–89. http://www.ohchr.org/EN/UDHR/Documents/60UDHR/detention_infonote_4.pdf. Accessed 26 July 2009.

United Nations. (2006). Convention on the rights of persons with disabilities. http://www.un.org/disabilities/documents/convention/convoptprot-e.pdf

Vale, C. L., Thompson, L. C., Murphy, C., Forcat, S., & Hanley, B. (2012). Involvement of consumers in studies run by the medical research council clinical trials unit: Results of a survey. *Trials, 13*, 9.

Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent

57

Norbert Konrad and Sabine Müller

Contents

Abstract

Medical treatments are allowed only with the informed consent of the patient. The German civil commitment laws contain articles, which allow the commitment of mentally ill persons, who are acutely dangerous for themselves and/or for important legal rights of third persons. However, neither a mental illness alone, nor missing insight into the illness and the need for treatment are sufficient for compulsory admission. The most controversial issue is whether persons who are unable to consent have a right to live with their illness, particularly their mental

N. Konrad (✉)
Institut für Forensische Psychiatrie, Charité – Universitätsmedizin, Berlin, Germany
e-mail: norbert.konrad@charite.de

S. Müller
Department for Psychiatry and Psychotherapy CCM, Charité – Universitätsmedizin, Berlin, Germany
e-mail: mueller.sabine@charite.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_56,

illness, or rather a right for an effective treatment of the illness (implicitly acknowledging the subjective torment and, at times, sheer terror the symptoms cause for the psychotic individual). In Germany, compulsory treatments of patients under custodianship and in confinement can be allowed by courts if certain conditions are fulfilled: The patient is not able to consent to the necessary treatment; the physicians have tried to convince him; the treatment is necessary to avert considerable health detriments; the compulsory treatment is used as a last resort treatment, and its benefit-risk balance is positive. A rather new instrument for avoiding compulsive measures in states of dangerousness is the mental health advance directive. Preliminary results are promising: Patients who have written advance treatment directives committed less violent acts, needed less use of social workers time, showed greater improvement in their working relationship with their clinicians, and were more likely to report satisfaction with their treatment.

Legal and Ethical Issues of Compulsory Interventions in Mentally Ill Persons

Generally, medical treatments are allowed only with the informed consent of the patient. Therefore, it is both ethically and legally unacceptable to treat a legally competent patient without his consent or even against his declared will. The right for autonomous decision making includes the right to die and therefore the right to refuse life-saving treatments. This conception has become predominant in medical ethics during the last decades; by now, it has entered the legislations of many countries. Particularly in Germany, the civil law (BGB § 1901a) clearly states that a legally competent adult can rule out certain life-saving treatments by declaring his will in an advance directive. The will of a mentally ill individual has to be respected, too, even if the patient refuses medically necessary treatments – provided that he or she is able to consent. Consent to treatment should be sought from all patients, including offenders suffering from a mental disorder, provided that they have the capacity to consent. If patients lack this capacity, physicians should seek their assent, even if the consent of the custodian would be legally sufficient, i.e., they should inform the patients about the intended treatment in adequate wording, so that the patients can understand the information – and ask them whether they assent to the treatment (Sammons 2009). Obtaining the patient's consent is essential to build up a "therapeutic alliance," which increases the likelihood of committing the patient to the treatment offered. This does not mean that psychiatrists should try to persuade patients to accept the proposed treatment; rather, it means that psychiatrists should try to convince patients with truthful information about the proposed treatment even though the patient might have refused it initially, be it because of fear, disinformation, or psychotic ideas. Each ethically acceptable effort should be made to convince the patient to cooperate. The use of coercion, allowing for "choosing" between the two evils of physical restraint and medication, should be avoided.

Nevertheless, a few patients will remain who will not consent to the proposed treatment and whose suffering will foreseeably increase or who will even die

without treatment. In these cases, physicians are confronted with the ethical dilemma whether the patients' right to decide autonomously prevails over their right for physical and mental health and for life. Although the predominant ethical and legal systems favor autonomy over the right to live, for many physicians, this dilemma will remain hardly endurable.

The situation is even more complicated for persons who are unable to consent – be it because of acute mental disorder (e.g., psychosis, suicidal ideation), mental disability, severe dementia, or coma. Nevertheless, also for these patients, decisions about any medical treatments have to be made according to their will: firstly, according to their formerly declared will (ideally in an advance directive), secondly, to their assumed will, and thirdly (in case that the latter two are unknown), in their best interest.

The most controversial issue is whether persons have a right to live with their illness, particularly their mental illness, or rather a right for an effective treatment of the illness (implicitly acknowledging the subjective torment and, at times, sheer terror the symptoms cause for the psychotic individual). The Federal Constitutional Court of Germany has supported the concept of the "freedom for illness," which means the right to refuse certain or all medical treatment and is part of the constitutional rights of patients for autonomous decision making. In a literal understanding, this right could imply that the society has to leave persons suffering from their illness if they refuse any treatment. This interpretation is favored as well by neoliberal opinion leaders as by anti-psychiatrist groups. Nevertheless, this interpretation is too narrow: As the medical ethicists Beauchamp and Childress have argued, the obligations to respect autonomy do not apply to persons who show a substantial lack of autonomy, because they are immature, incapacitated, ignorant, coerced, or exploited, for example, infants, irrationally suicidal individuals, severely demented subjects, or drug-dependent patients (Beauchamp and Childress 2013, p. 108). This position has also recently been supported by the Federal Constitutional Court of Germany (in accordance with the Convention on the Rights of Persons with Disabilities) in a decision about forced treatment of a forensic patient: The freedom for illness must not be considered detached from the real capacities of free decision making which may be limited by illness. Therefore, the state is not obligated to leave forensic patients to the fate of permanent confinement because of the primacy of an illness-determined will, but coercion has to follow a rule of law (Bundesverfassungsgericht, 2 BvR 882/09, 52, 23.03.2011). Therefore, the state parliaments are urged to reformulate their civil commitment laws so that compulsory treatments will be made possible under strict legal conditions.

Nevertheless, the aforementioned decision has an immense influence on jurisdiction about involuntary treatment and has strengthened the position that the patient's autonomy dominates over the patient's right for mental health or for life, and overrules his best interest as seen by independent custodians or medical professionals. Particularly, the Federal Constitutional Court has judged that coerced treatment against the will of the forensic patient is not allowed to protect other persons against criminal acts which the patient might commit after discharge. The Court argues that future crimes could be prevented by detaining the patient in a psychiatric institution without treatment.

Following the aforementioned decision of the Federal Constitutional Court of Germany, several local courts and district courts have applied this decision to lawsuits about the issue of compulsory treatment of mentally ill patients under legal custodianship. They decided that the German civil law does not allow compulsory treatment. This interpretation was supported by the Federal Supreme Court (Bundesgerichtshof, 20.06.2012, XII ZB 99/12) which gave up its former position in order to follow the decision of the Federal Constitutional Court of Germany. Within only 7 months, the relevant articles of the German civil law have been modified. Since January 2013, compulsory treatments of patients under custodianship and in confinement can be allowed by courts if certain conditions are fulfilled: The patient is not able to consent to the necessary treatment; the physicians have tried to convince him; the treatment is necessary to avert considerable health detriments; the compulsory treatment is used as a last resort treatment and its benefit-risk-balance is positive. This law can also be applied to forensic patients under custodianship.

Although the legality of commitment and of compulsory treatment depends completely on the label "capacity to consent," there is no standardized and validated tool to assess this capacity. Indeed, there are at least three instruments trying to measure the ability for informed consent, namely, the Mac Arthur Competence Assessment Tool-Treatment (Mac CAT-T) (Grisso and Appelbaum 1998), the Hopkins Competence Assessment Test (HCAT) (Janofsky et al. 1992), and the assessment tool of the AGFP (Nedopil et al. 1999). These instruments try to quantify a patient's ability to understand, appreciate, reason about, and express a therapy choice. It has been rightly criticized that the Mac CAT-T is merely orientated on the rational understanding, but neglects emotional and value-based prerequisites of decisions which are often affected by psychiatric disorders (Vollmann 2008), a shortcoming, which the AGFP instrument tries to avoid. Furthermore, it has been criticized that the neurological prerequisites of autonomous decision making are neglected even in the medical ethical debate about autonomy (Müller and Walter 2010). In medical practice, the capacity to consent is evaluated often rather informally or based on the personal psychiatric experience. Surprisingly, the Federal Constitutional Court has not criticized this shortcoming (Müller et al. 2012a), seemingly based on its general assumption that professionals are free in the choice of medical methods if they adhere to the current state of the art.

In Germany, compulsory treatments are legal in situations of imminent danger for self or others, if certain conditions are fulfilled.

Legal and Organizational Frame for Mentally Ill Offenders

Compulsory Admission

In Germany, mentally disordered offenders are subject to special legal regulations (Konrad 2001), which are based on the concept of criminal responsibility: Offenders who are not criminally responsible and not considered dangerous are hospitalized, if at all, in general clinical psychiatric institutions. If serious

offenses are expected from offenders who are considered to have at least diminished criminal responsibility, they are admitted, regardless of therapeutic prospects, to special forensic psychiatric hospitals (Art. 63 German Penal Code) under the authority of the state ministries of health. The number of detainees housed there was 6,750 as of March 31, 2012 (www.destatis.de). Offenders dependent on psychoactive substances with sufficiently good therapeutic prospects, independent of their assessment of responsibility, are admitted to special forensic drug treatment facilities, which are also under the authority of the ministry of health (Art. 64 German Penal Code). As of March 31, 2012, the number of detainees housed there was 3,526 (www.destatis.de). All other mentally disordered offenders, including individuals with schizophrenia who are considered criminally responsible despite their illness, may be sentenced to prison, if no milder sanctions like a fine are ordered by the court. In many cases, it may depend on coincidental constellations whether a mentally ill offender is committed to a forensic psychiatric or penal institution (Marneros et al. 2002).

Prisoners who pose a danger to themselves, for example, after a suicide attempt or other self-destructive behavior, are frequently admitted to psychiatric wards within the prison. There are, however, no legal regulations about the admission, treatment, and dismissal of those patients. The penal detention code and the federal and state laws neither stipulate nor prohibit psychiatric prison wards.

Compulsory Treatment and Force-Feeding in Prisons

If mentally disordered prisoners are under legal custodianship, the custodian can request a medically indicated compulsory treatment from the court according to the new civil law. For all other mentally disordered prisoners, compulsory treatment is regulated by the penal law, the pertinent provisions of which correspond to the standards for compulsory treatment within the framework of state commitment laws (PsychKG). Compulsory treatment occurs within psychiatric facilities of prison hospitals. Unlike Sweden, German rules do not make it necessary to send the prisoners to general psychiatric facilities if compulsory measures become necessary (Salize et al. 2007).

Occasionally, the patient's decision to refuse treatment results from a conflict relating to nonmedical issues: for example, when a prisoner goes on hunger strike to protest against a judicial or administrative decision. In this situation, the doctor should assess the state of health of the person concerned and subsequently make a detailed note in the patient's file to document that the individual has the capacity to understand the treatment proposed but has refused treatment on sound intentions after being given detailed information. Psychiatrists are regularly asked to assess the mental state of prisoners, especially to answer the question if the refusal is caused by delusions (e.g., to be poisoned).

The need for medical care of prisoners who persistently refuse food in order to make a protest is rare but challenging. Knowledge about the hunger strike quickly

spreads and gets into the political arena. Governments want to resist the demands, which often have political implications, but do not want prisoners to die. Pressure is therefore imposed on the prison health care staff, including psychiatrists, to keep the prisoners alive, if necessary, by force-feeding. However, a doctor must obtain consent from the patient before treating him. The only exception is an emergency when the patient is incapable of giving consent. Since the end stage of food refusal is coma, it is foreseeable that the patient will become incapable of giving consent. At that stage, doctors are allowed for intervening by artificial feeding to save the patient's life. However, this is not allowed if the patient has made it clear beforehand that he refuses interventions to prevent death (e.g., by an advance directive).

Disciplinary Measures

Mental health problems may be overlooked especially in prisoners who are psychotic quietly. The more behaviorally disturbed are often viewed as a disciplinary problem rather than as individuals with mental health needs (Birmingham 2004). Some of them are placed in disciplinary segregation instead of immediately receiving appropriate psychiatric care. Of particular concern are disciplinary measures which are coercive by nature. Mentally disordered prisoners are more likely to become the subject of disciplinary measures due to their misbehavior that may be caused by the disorder (Birmingham 2004; Morgan et al. 1993). It is well known that specific coercive measures (e.g., solitary confinement) are likely to aggravate mental disorders. Thus, it is crucial to assess the mental state of a prisoner prior to implementing such measures in order to avoid any additional harm. In some European countries, e.g., Germany, all prisoners for whom punitive or disciplinary measures are intended or all prisoners known to suffer from a mental disorder are assessed for their resilience before disciplinary measures are executed. In other European countries, such an assessment is not stipulated (Salize et al. 2007).

The participation of medical personnel in the administration of disciplinary measures raises considerable ethical problems: Discipline and punishment are security but not health issues, and therefore, the physician, who should be available to attend to the medical needs of prisoners under any form of punishment, has no role in deciding about the administration of such punishment, e.g., in certifying that a person is mentally fit to withstand such a punishment, and should not be available for the purpose of supporting the prisoner's capacity to sustain a punishment (WHO Europe 2009).

Somewhat surprisingly, in Germany, like in most European countries, disciplinary or coercive measures during imprisonment must be recorded but are not published, so that scientific analyses are not possible. Such records or files would be an essential tool for investigating the appropriateness of such measures, particularly in the case of mentally disordered prisoners. Because of the negative effects of disciplinary measures on the mental health – especially for mentally disordered patients – close confinement should be reduced to an absolute minimum and be replaced with one-to-one continuous nursing care as soon as possible. Unfortunately, in Germany, there is more isolation and observation by video than

one-to-one continuous nursing. In such cases, the prison psychiatrist is often confronted with ethical conflicts: Testifying acute suicidality in a mentally disordered prisoner without the possibility of adequate inpatient treatment means to produce a possibly traumatizing situation, particularly a situation of isolation with the requirement to undress or change clothes and to being exposed to video observation.

Legal and Organizational Frame for Dangerous, Mentally Ill, Nondelinquent Persons

Also nondelinquent persons with mental illness can be admitted to psychiatric institutions and treated there against their will, if they are acutely dangerous for themselves or for third persons. The civil commitment laws of the federal states of Germany (PsychKG or Unterbringungsgesetze) rule the compulsory admission and the compulsory treatment of this group of patients. Although each of the federal states of Germany has its own civil commitment law, they coincide in their core issues and differ only in some details.

Compulsory Admission to General Psychiatric Clinics

The civil commitment laws contain articles, which allow the commitment of mentally ill persons, who are acutely dangerous for themselves and/or for important legal rights of third persons. However, neither a mental illness alone, nor missing insight into the illness and the need for treatment are sufficient for compulsory admission. A compulsory admission has to be ordered by a court.

Compulsory Treatment

The civil commitment laws also allow for compulsory treatment of patients admitted to a psychiatric institution, but, as mentioned above, the Federal Constitutional Court has recently judged these regulations of two federal states as unconstitutional, and prompted the state parliaments to reformulate their laws so that compulsory treatments could be allowed, if they regarded the strict legal conditions, which have been formulated by the Federal Constitutional Court (Müller et al. 2012a). Several federal states are currently working on reforms of their civil commitment laws.

Only in cases of acute danger (e.g., if a patient tries to hurt or even kill someone), it is allowed to use compulsory treatments (Penal Code, Art. 34). Nevertheless, in most such cases, other measures will be preferable, e.g., fixation or isolation.

Compulsory treatments of confined mentally ill patients under custodianship can be allowed by courts according to the modified civil law. Therefore, a number of conditions have to be fulfilled: The patient is not able to consent to the necessary treatment; the physicians have tried to convince him; the treatment is necessary to avert considerable health detriments; the compulsory treatment is used as a last

resort treatment and its benefit-risk balance is positive. Compulsory treatments according to civil law can be allowed only for the patient's own benefit; they are no means to prevent harm from third persons.

Further Compulsory Measures

Besides compulsory medical treatments, a number of nonmedical measures are used either alternatively or additionally to compulsory medical treatment, in particular isolation, separation, fixation, mechanical restraint, and physical restraint. These measures are regulated in the civil commitment laws; they have not been questioned by the Federal Constitutional Court.

Measures to Avoid Compulsion

A rather new instrument for avoiding compulsive measures in states of dangerousness is the mental health advance directive. Advance directives are nowadays established for end-of-life settings, but they might also be suited to mental health settings. A patient with a chronic mental illness can specify her future preferences for treatment, should she lose the capability to make decisions. By way of example, patients can specify whether they prefer isolation, mechanical fixation, or medication, and where required, which drugs, the maximal dose of drugs, and the maximal period of drug application. According to a Cochrane study (Campbell and Kisely 2010), there are too few data available to make definite recommendations, but the preliminary results are promising: Patients given advance treatment directives committed less violent acts, needed less use of social worker's time, and showed greater improvement in their working relationship with their clinicians and were more likely to report satisfaction with their treatment. High intensity advance directives, such as joint crisis planning, may offer promise.

Conclusion and Future Directions

Generally, a patient-centered treatment which is orientated to the patients' needs and avoids restrictions as far as possible allows for reducing compulsory treatments to the minimum.

The mental health advance directive seems to be useful to avoid compulsive measures in states of dangerousness. Furthermore, if "permanently closed" wards are abolished and the doors of all treatment units are opened as often as possible, aggressive incidents, compulsory treatment but also absconding can be substantially reduced (Lang et al. 2010).

Acknowledgment The authors thank Prof. Dr. Norbert Nedopil for his critical comments on an earlier version.

Cross-References

▶ Ethics in Psychiatry
▶ Strengthening Self-Determination of Persons with Mental Illness

References

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). Oxford: Oxford University Press.

Birmingham, L. (2004). Mental disorder and prisons. *Psychiatric Bulletin, 28*, 393–397.

Campbell, L. A., & Kisely, S. R. (2010). Advance treatment directives for people with severe mental illness. *The Cochrane Collaboration. The Cochrane Library*, 1–38.

WHO Europe. (2009). *Trenčín Statement on prisons and mental health*. Copenhagen: WHO Europe, http://www.euro.who.int./Document/E914202.pdf. Retrieved 10 Nov 2012.

Grisso, T., & Appelbaum, P. S. (1998). *Assessing competence to consent in treatment. A guide for physicians and other health professionals*. New York/Oxford: Oxford University Press.

Janofsky, J. S., McCarthy, R. J., & Folstein, M. F. (1992). The Hopkins competency assessment test: A brief method for evaluating patients' capacity to give informed consent. *Hospital & Community Psychiatry, 43*(2), 132–136.

Konrad, N. (2001). Redevelopment of forensic-psychiatric institutions in former East Germany. *International Journal of Law and Psychiatry, 24*, 509–526.

Lang, U. E., Hartmann, S., Schulz-Hartmann, S., et al. (2010). Do locked doors in psychiatric hospitals prevent patients from absconding? *The European Journal of Psychiatry, 24*(4), 199–204.

Marneros, A., Ullrich, S., & Rössner, D. (2002). *Angeklagte Straftäter. Das Dilemma der Begutachtung*. Baden-Baden: Nomos.

Morgan, D. W., Edwards, A. C., & Faulkner, L. R. (1993). The adaptation to prison by individuals with schizophrenia. *The Bulletin of the American Academy of Psychiatry and the Law, 21*, 427–433.

Müller, S., & Walter, H. (2010). Reviewing autonomy. Implications of the neurosciences and the free will debate for the principle of respect for the patient's autonomy. *Cambridge Quarterly of Healthcare Ethics, 2*, 205–217.

Müller, S., Walter, H., Kunze, H., Konrad, N., & Heinz, A. (2012a). Zwangsbehandlungen unter Rechtsunsicherheit. Teil 1: Die aktuelle Rechtslage zu Zwangsbehandlungen einwilligungsunfähiger Patienten mit psychischen Erkrankungen. *Nervenarzt, 83*, 1142–1149.

Müller, S., Walter, H., Kunze, H., Konrad, N., & Heinz, A. (2012b). Zwangsbehandlungen unter Rechtsunsicherheit. Teil 2: Folgen der Rechtsunsicherheit in der klinischen Praxis – Vorschläge zur Verbesserung. *Nervenarzt, 83*, 1150–1155.

Nedopil, N., Aldenhoff, J., Amelung, K., Eich, F. X., Fritze, J., Gastpar, M., et al. (1999). Competence to give informed consent to clinical studies. Statement by the taskforce on "ethical and legal questions" of the Association for Neuropsychopharmacology and Pharmacopsychiatry ("Arbeitsgemeinschaft für Neuropsychopharmakologie und Pharmakopsychiatrie [AGNP]"). *Pharmacopsychiatry, 32*(5), 165–168.

Salize, H. J., Dreßing, H., & Kief, C. (2007). *Mentally disordered persons in European prison systems – needs, programmes and outcomes* (EUPRIS). Final report. Mannheim: Central Institute of Mental Health.

Sammons, H. (2009). Ethical issues of clinical trials in children: A European perspective. *Archives of Diseases in Childhood, 94*, 474–477.

Vollmann, J. (2008). *Self-determination of patients and ability for self-determination (Patientenselbstbestimmung und Selbstbestimmungsfähigkeit)*. Stuttgart: Kohlhammer.

Relationship of Benefits to Risks in Psychiatric Research Interventions

58

Hanfried Helmchen

Contents

Abstract

The risk-benefit evaluation of a research intervention is only probabilistically possible and is open for contextual influences, because the criteria of benefits and of risks are often only insufficiently quantitatively defined. The question remains whether it is at all possible and, if so, how individual benefits and risks can be balanced against societal benefits and risks. Algorithmic attempts to structure the evaluation process should standardize the evaluation. However, for now, only a pragmatic solution will validate the result in three steps (researcher, ethics committee, potential research participant).

Introduction

Every physician must consider the expected benefits in relation to the potential risks of his/her diagnostic, therapeutic, prophylactic, etc., intervention

This paper is based on Chap. 3 Ethische Grundvoraussetzungen klinischer Forschung, 3.1 Nutzen und Risiken in Helmchen H (ed) Ethik psychiatrischer Forschung (2013) Springer, Heidelberg

H. Helmchen
Department of Psychiatry & Psychotherapy, Charité – University Medicine Berlin, CBF, Berlin, Germany
e-mail: hanfried.helmchen@charite.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_60,

for the patient. This up to now mostly implicit estimation has become more explicit during the past decades by the medical obligation to inform the patient in order to gain his/her consent. This is particularly valid in the case of off-label use of interventions or quasi-experimental procedures in the individual patient, the so-called attempts at healing (Heilversuche). Moreover, especially in research interventions, the researcher has to consider explicitly his/her arguments for the acceptability, i.e., the justification and reasoning, of his/her planned research application to an ethics committee (Europarat 2005). Several regulations specify that the potential risk must have been assessed adequately and can be dealt with satisfactorily (World Medical Association 2008).

Without these prerequisites, a research intervention is not permissible even if a subject with the capacity to consent consents to participate. However, this does not mean that risky interventions or those without potential individual benefit cannot be justified ethically in consenting subjects, e.g., healthy volunteers in phase 1 trials. Thus, reasons must be given for an acceptable benefit-risk relationship.

But it is difficult to find an acceptable benefit-risk relationship,[1] or it will be seen as impossible: “risk-benefit ratios often cannot be calculated, even roughly; and that even if they could, ethical experiments don’t need to have favorable risk-benefit ratios” (Rajczi 2004). The final report of the American National Bioethics Advisory Commission reads as follows: “An IRB may approve a research proposal only if it judges that the risks are reasonable in relation to potential benefits. This judgment may be an IRB’s single most important and difficult determination, because it ensures that when research participants voluntarily consent to participate in a research study, they are offered a ‘reasonable choice’” (cit. Simonsen 2009). Unfortunately, as the report continues, “current regulations do not further elaborate how risks and potential benefits are to be assessed, and little additional guidance is available to IRBs” (Wendler and Miller 2007).

This has to do primarily with the estimation of benefits to risks in the participating individual. But it is also an estimation of the individual benefits and risks to those for society. However, there is a doubt that it is possible at all to estimate the relationship of these individual benefits and risks to the potential benefits and risks for society other than qualitatively and personally. But there is “not at all an operational criterion for the decision that this benefit or harm for the individual is of greater magnitude than benefit or harm for society. Furthermore, there is no way to calculate the social value against the individual risk without further assumptions.” (Wiesing 2011). Therefore, Hüppe and Raspe avoid terms which may “suggest a comparability of the extreme heterogeneous potentials of benefits and risks” (Hüppe and Raspe 2011).

[1] “The wording ‘fair balance’ is occasionally used by the European Court of Human Rights when there is a reasonable relationship between legitimate but conflicting interests, typically between the individual and the society at large.” (Simonsen 2009).

Example 1
Compare the potential societal benefit of sequencing the human genome for possible targets of therapeutic interventions with the (minimal) individual risk of participants for discrimination and stigmatization by a misuse of their individual genetic data. Ethically it seems important that on the one hand such research interventions are without potential individual benefit and, furthermore, the confidentiality of the individual genetic data cannot be guaranteed with certain (Maier et al. 2013), and on the other hand, it is not for sure that the modern sequencing methods will reach their promised objectives.

Example 2
The individual benefit of recovery from an illness as quickly as possible may clash with the social value of gain of knowledge if, e.g., the patient's recovery may be delayed because he/she belongs to a purely placebo group.

In these examples, the gain of general knowledge for the benefit of future patients as members of society can be understood as a social benefit (Emanuel et al. 2000).

Raspe relates the term "societal benefit" in a much broader sense to other "beneficiaries: the whole explanatory knowledge of medicine, population health, health insurance and their financial stability..." (Raspe 2012). However, to judge the relationship of such benefits to risks may be much more difficult than that between the defined benefits and risks of a specific medical intervention both for the individual research participant and the societal benefits and risks.

Furthermore, a basic difficulty exists insofar as potential risks and benefits can be determined only as probabilities, e.g., as "probable," "possible," or "not to be excluded." Moreover, these probabilities vary among individuals, e.g., with regard to every day risks, or among ethics committees. Due to the fact that unequivocal criteria of the extent of benefits and risks as well as clear algorithms for the evaluation of their mutual relationship do not exist (Rid et al. 2010), the justification of a benefit-risk relationship can be influenced by individual dispositions (characters or prejudice) as well as by the current social situation of the evaluator, the physician, the patient, the potential research participant, and members of the ethics committee (Rid et al. 2010). At least the evaluator should be aware of this and reflect such possible background influences on his/her preconceptions: e.g., a society-oriented evaluator (or the researcher) could estimate the societal benefit of the expected gain of knowledge more highly and the potential burdens of the proband less than a more individualistic or research skeptical evaluator (or treating physician) might do.

Rid et al. state six reasons for the unreliability of such intuitive judgments among others that they do not consider systematically existing empirical data and that they are subject to personal bias, e.g., by rating lower the risks of interventions familiar to

the evaluator than the risks of interventions with which he/she is not familiar, and that therefore the estimates of acceptable risks vary in a broad range among ethics committees (Rid et al. 2010; Shah et al. 2004).

Because of these difficulties of judgment, research ethics committees tend to avoid comprehensive evaluations of the risk-benefit relationship and focus on other aspects of the research project, such as the informed consent process, as Simonsen found in a 3-year observational study of Norwegian research ethics committees (Simonsen 2012).[2]

However, because researchers and ethics committees must estimate the benefit-risk relationship of a research project in order to account for legal regulations, they should communicate the reasons for their estimates with regard to their comprehensibility. And, where applicable, they should say that "defined risks are not acceptable, namely in the sense that they are not negotiable" (Wiesing 2011). In any case it is the task of the researcher to convey the significance of probabilities and particularly those of the benefit-risk estimation in a mode that can also be understood by the potential research participant. In view of the uncertainty of risks that can be calculated only as probabilities, decisions will be made logically statistically according to the prevailing opinion, but in fact however intuitively heuristically (Gigerenzer 2006). The question of whether such knowledge of the psychology of intuitive decisions will be helpful requires future investigations.

The evaluation of the acceptability of the benefit-risk relationship is specifically important in research interventions with patients whose capacity to consent is impaired by a mental illness, because the risk of exploitation of such vulnerable subjects may be larger than in patients who are competent to give informed consent. A careful evaluation also includes an understanding of the uncertainties in assessing potential benefits and risks that should be considered for both the individuals participating in research and other current or future patients (society).

Thus, benefits and risks are to be evaluated primarily with regard to individual subjects who participate in a research intervention. In cases with more than minimal risks on the individual level benefits and risks are to be considered with regard to society too.

> Also interactions have to be considered among individual and societal benefits and risks. This is valid not only for research interventions but also for everyday medical interventions. Thus, societal risks such as burdens on insurance companies should also be considered on the individual level, because, in order to save resources for the community, medical services may be applied only in an effective and economically efficient mode and their benefit must be proven according to the state of the art (SGB V) (Deutscher Bundestag 2010).

In order to make clear the determinants of such estimations, Rid et al. recently proposed a procedure for a standardized evaluation of risks, which will be considered in the following (Rid et al. 2010).

[2]Literature on the lack of rules and on the difficulties of benefit-risk estimates of research projects can be found in (Hüppe and Raspe 2011) and in (Rid et al. 2010).

Type, Magnitude, and Likelihood of Benefits and Risks

The comparability of benefits and risks as well as the evaluation of the benefit-risk relationship requires a:
- Definition of specific types
- Graduation (or even quantification)
- Estimation of the probability of occurrence of benefits and risks

Benefits

Individual Versus Social Benefit

The *social benefit* of clinical research in psychiatry consists in the gain of knowledge for the improvement of the treatment and care of mentally ill patients. Basically it can be seen as urgently desirable, because mental diseases are widespread and the need for research is great as opposed to the necessary scientifically proven knowledge for optimal psychiatric action. However, the estimation in the individual case depends:
- On the one hand upon the probability of an unequivocal gain of knowledge, i.e., upon the quality of the scientific method
- But on the other hand upon the relevance of the expected gain of knowledge for psychiatric action

Due to the conviction of liberal western societies expressed legally and also in the Declaration of Helsinki, § 6 (World Medical Association 2008) that no human being is obliged to donate himself/herself to society, the practice of clinical research will be dominated less by its societal benefit than by the individual benefit of the patients participating in research.

Individual benefit comprises both:
- The subjectively determined well-being in the sense of a self-experienced and self-evaluated benefit
- The more intersubjectively (“objectively”) evaluated “best interest” of the patient as a benefit seen from the outside, so to speak an objectively, externally judged benefit

This differentiation is significant insofar as in case of a patient not competent to consent, the substituted consent by an authorized person must be guided by wishes of the patient expressed beforehand, e.g., by an advance directive, for the well-being of the patient. If no information exists about such personal wishes, the substitute must decide in the best interest of the patient (Heinrichs 2007).

Patients themselves see the benefit of participating in research in that they will:
- Receive a better treatment that will be more effective than the available standard therapy or will act more quickly or will have fewer side effects
- Satisfy altruistic feelings of solidarity with other similarly ill patients (Rosenbaum 2012)

> Most respondents continue to participate in the ESPRIT study in hopes of benefiting personally. The majority also recognized that by participating in ESPRIT they were contributing to helping others; they experienced pride regarding this contribution and considered it an important reason to continue to participate. (Magnus and Merkel 2007; Wendler et al. 2008)

- Receive money or other advantages (Sofaer et al. 2007)

Further motivational factors are:
- To receive more information about one's own illness and its characteristics.
- To feel self-determined.
- The hope that other people will understand better their mental state.
- Particularly in the mentally ill without the capacity to consent, the motivation of the caretakers is important; this has been evidenced in research interventions which have aimed for an improvement of the quality of life of the patients and/or for an attenuation of the burdens of the caretakers (Connell et al. 2001; Mastwyk et al. 2002).

Specifications of Benefit

Benefit can be defined only in relation to something:
- *Societal* benefit of research is related to the gain of knowledge. In this case, the "essentiality" of the gain of new knowledge towards the existing knowledge plays an important role (s. below).
- *Individual* benefit, e.g., can be assessed as an attenuation of symptoms or of suffering or of an augmentation of the quality of life or functional improvement.

Compared with clearly defined and well-ascertainable phenomena such as some symptoms, it is more difficult to operationalize the attenuation or increase of more complex phenomena such as suffering or quality of life; but such an operationalizing would be a prerequisite or at least a support for the assessment of the size of the benefit. However, many of the following terms, which specify and grade benefits and risks by the dimensions of size and probability, are not clearly defined or are not at all definable and thus are open to subjective interpretations.

Such specifying criteria of benefit are:
- "Direct" or "immediate" benefit will be used synonymously. However, "direct" benefit can be understood as a causal effect of the intervention, "immediate" benefit by contrast as an effect in a timely connection. "Direct" suggests that "indirect" forms of benefit may also exist, e.g., if the development of a new effective therapy is based on the knowledge of the cause of the illness that was found during an earlier research intervention in a patient with an illness of long duration. "Few existing accounts disagree over how this crucial concept of 'direct' benefit should be defined. This disagreement raises concern over whether those who cannot consent, including children and adults with severe dementia, are being adequately protected." It is suggested "that the extant definitions of direct benefits either provide insufficient protection for research subjects or pose excessive obstacles to appropriate research" (Friedman et al. 2010, p. 60).

- "Therapeutic research" as potentially beneficial for the patient participating in research had been compared to "nontherapeutic research" without potential individual benefit. However, this distinction is questionable, because the border between both types of research is often not clear: "A therapeutic research study may prove that the experimental intervention is ineffective, in which case undergoing the experimental condition would be not beneficial to the subjects. Conversely, a non-therapeutic study may be associated with benefits for the subjects, such as more attention from health care workers. etc." (Welie and Berghmans 2006, p. 69). This is particularly valid with regard to the "therapeutic misconception," i.e., that the research participant misunderstands the research intervention as a mere therapeutic intervention (Vollmann 2000). Therefore, the ethically unequivocal terms "with" or "without" potential individual benefit should be preferred.
- As "collateral" benefit, a benefit will be termed that cannot be ascribed causally to the research intervention but to other aspects of the study performance and study participation, e.g., an "inclusion benefit" by an intensive medical monitoring (Hüppe and Raspe 2011).
- "Important," "essential," or "significant" benefit are particularly vague terms and are thereby open for different interpretations. Thus, the Explanatory Report (Nr. 87) to the Additional Protocol of the European Council defines "essential" as an "essential extension of the scientific understanding of a disease" (Europarat 2011). The circularity of this explanation shows the difficulty in defining the term "essential" clearly, unequivocally, and practically. Aside from a broad understanding of the contents of the field in which "essential" insights can be gained, i.e., new knowledge of causes, treatment, and prevention of a disease, the term "essential" itself remains unclear.

  Is it necessary for new knowledge, in order to be viewed as "essential":
  - To be not less than a breakthrough, i.e., knowledge that opens new possibilities for action?
  - To be a breakthrough only with an immediate impact or also with a delayed effect?
  - To be – with regard to formal criteria – proven at least and on which level?

    On the other side, the vagueness of such terms opens a necessary range for interpretations, because the newness of knowledge progress and its practical usefulness are difficult to judge and are only seldom quickly recognizable.
- Because it is the objective of research to gain new knowledge, every research surpasses an exclusively individual benefit and is oriented supraindividually. Insofar the extent of the individual benefit, i.e., the self- interest, must be related to the benefit for others, i.e., other person's interests ("Fremdnützigkeit"). Hence, the potential individual benefit (benefit, potential individual) is largest if the research participant can expect such benefit for himself/herself during the ongoing study (group 1 of the statement of the Central Ethics Committee of the German Board of Physicians (Zentrale Ethikkommission bei der Bundesärztekammer 1997)); it will be less in studies with only future benefit (group 2) and

will be at most questionable in research with expected benefit mainly for the group of patients to which the research participant belongs by age or disease (group 3). Such "group-specific" research will be differentiated from research with benefit exclusively for others (group 4).

In the framework of clinical research "group-specific" benefit as benefit for others means improved medical knowledge for the optimization of diagnostics, therapy, or care for other patients with the same disease or in the same age group. This can be the sole benefit of a research intervention with only questionable or no potential individual benefit for the participant of a research intervention, e.g., in a validation of a diagnostic measure or in a study of possible conditional or risk factors or causes of a disease. In order to prevent an excessive use of the term "group benefit," it has been proposed to define further beneficiaries not only by disease, age, and gender but additionally by the "inclusion and exclusion criteria on which the study is based" (Hüppe and Raspe 2011).

- The "size," "extent," "magnitude," or "strength" of a benefit however defined could be graded as "questionable," "slight or recognizable," or "unequivocal or strong."
- "Prospective" or "potential" benefit indicates an anticipation or expectation of a benefit. Because it is a determination of a likelihood, the occurrence of a benefit should at least be graded as "possible" or "probable."

Risks

If a subject participates in a necessary or even legally required research intervention for the benefit of all, he/she must be protected against the risks of this intervention. Protective norms are also related to gradings of risks.

Various normative rules describe the content, extent, and the mode of protection of research participants against risks, e.g., declarations such as the best known one of Helsinki in 1964 with the following revisions by the World Medical Association (World Medical Association 2008), national research laws, and particularly the first internationally binding instrument for biomedical research, the European Convention on Human Rights and Biomedicine (1997) (Europarat 1997) and its Additional Protocol (2005) (Europarat 2005).

Individual Versus Societal Risks

Individual Risks (*Risks, Individual*)

The term risk comprises both:

- Rather "objective" endangering of the individual research participant, e. g., with unwanted accompanying effects of the intervention, for example, unwanted side effects of an investigational drug or risks of blood taking for research purposes, but also prolongation of suffering or worsening of the disease due to withholding a specific treatment in a placebo group, in a broader understanding also

dispositions for unwanted effects such as (pharmaco)genetic or allergic inclinations belong to this group of risks.
- Rather "subjective" burdens and inconveniences, e.g., by the methodological rigor of the research procedure or feared risks of stigmatization especially in depressive or drug-dependent patients, which may demotivate potential research participants. Because of the subjective dimension of risks and unwanted effects, potential research participants should be examined particularly for their susceptibility to physical or psychic burdens that may be related specifically to the research intervention.

  Thus, e.g., the magnetic resonance imaging (MRI) has no objective risks, but may indeed – in claustrophobic patients – become a subjective burden that leads to a termination of the intervention. For risks that are specific for the intervention and that have large relevance for the further life of the patient, he or she must be informed independently of the probability of occurrence.

Societal Risks (*Risks, Societal*)

However, societal risks should also be considered, e.g., if research interventions have considerable risks or do not follow precisely the normative prescriptions and thereby lead to incidents that undermine the necessary trust of the public or of potential research participants. This can protract or even prevent the recruitment of research participants. But not executing a research project may also present a societal risk, e.g., if the lack of knowledge must urgently be eliminated, in order to contain and treat an acute threat to health such as an infectious epidemic or to substitute an ineffective measure by an effective one.

Specifications of Risks (*Risks, Specifications of*)

- Magnitude, extent, or strength of risks are graded by a broad range of terms, such as "without the danger of injury," "minimal risk," "minor increase of minimal risk," "not negligible risks," "serious risk for health," "potentially irreversible damage," and "risk of inacceptable dimension" (Helmchen 2002).
- Absolute upper limits of risks are irreversible impairments and death. Standard upper limits for research with patients incompetent to give informed consent are "no more than minimal risk" or with minors even "minor increase of minimal risk" (Wendler 2008).
- "No more than minimal risk" is a decisively limiting criterion for research with patients incompetent to give informed consent. However, there are different interpretations of "minimal risk":
  – US regulations permit ethics committees to approve a research intervention in patients incompetent to give informed consent only if "it poses no more than 'minimal' risk, defined as the risks encountered in daily life or during the performance of routine examinations or tests" (Wendler 2008, p. 467). But "in the absence of empirical data, IRB members may assume they are familiar with the risks of daily life and with the risks of routine examinations and tests and rely on their own intuitive judgment to make these assessments.

Yet intuitive judgment of risk is subject to systematic errors, highlighting the need for empirical data to guide IRB review and approval of pediatric research...Current data on the risk of mortality in healthy children suggest IRBs are implementing the federal minimal risk standard too cautiously in many cases" (Wendler et al. 2005). On the other hand, this concern has led to a warning against a softening of the "minimal risk" concern (Gefenas 2007).

– Furthermore, standards of minimal risk vary with regard to the risks of everyday life with age (Wendler 2009). With children there are other standards than with older persons. Because of these difficulties, it has been proposed that the standard of everyday life be eliminated (Resnik 2005).
– However, there has been an attempt to objectify the standard of everyday life by considering the magnitude and likelihood of injuries in sports and in car driving, i.e., two common everyday activities (Rid et al. 2010).

> With regard to the criterion "routine examinations," the Central Ethics Committee of the German Federal Board of Physicians commented that the standard of minimal risk is present "e.g., when taking small amounts of body fluids or tissue in the framework of diagnostic measures or surgery without additional risks for the patient. Defined somatic investigations also belong to this standard (e.g., sonography, transcutaneous tissue measurements) or psychological investigations (e.g., interviews with questionnaires, tests, observations of behavior)" (Zentrale Ethikkommission bei der Bundesärztekammer 1997).

- "Minor increase above minimal risk": this standard was introduced in the USA for research with children. However, it remains unclear what "minor increase" means (Wendler and Emanuel 2005). Accordingly other countries did not follow this line. Also the ambiguity of this criterion has led ethics committees to different interpretations and to a call "for a national consensus on the interpretation of federal regulations" (Fisher et al. 2007).
- Research without potential individual benefit in patients incompetent to give informed consent will be seen either – according to German law – as inadmissible or – according to the European Convention on Human Rights and Biomedicine (1997) – as ethically acceptable only as an exception and when limited by the criterion "no more than minimal risk," if at least a group-specific benefit can be expected and if the consent of the research participant will be substituted by an authorized person.

Proposals for Grading

During the past 15 years, the grading of benefits and risks has been intensified more systematically:

- Thus, a taxonomy of benefits and risks according to their magnitude or severity has been proposed (Helmchen et al. 1995) (see Table 58.1).
- The following table had been developed as an advanced categorization that contains the essential variables of evaluation of the German Drug Law (Deutscher Bundestag 2004) (see Table 58.2).

Table 58.1 Taxonomy of benefits and risks according to their magnitude (Translated in English and modified from Helmchen and Lauter 1995, S. 47f.)

Risks
1 No or at most minimal risk
2 Mild increase of minimal risk
3 Unequivocal more than minimal risk
4 Risk with irreversible consequences
Benefits
1 No or at best questionable benefit
2 Benefit only for the public good (benefit *without* potential individual benefit)
a. Only by extension or safeguarding of existing knowledge
b. By qualitative new knowledge
3 Benefit for both the public good and the individual research participant (intervention *with* potential individual benefit)
a. Only by quantitative improvement of existing standards
b. By qualitative innovative treatments

Table 58.2 Grading of risks and the essential evaluation variables of the German Drug Law (AMG) (Translated in English from Terwey (2007), S. 138, Table 24)

	Risk	Vulnerability	Chance	Scientific quality GRADE	Evidence SIGN
A	No risk	Not vulnerable	Patient	High	A
B	Minimal risk	Vulnerable	Group specific	Moderate	B
C	Minimal increase above minimal risk	Children	Science Short termed	Low	C
D	More than minimal increase above minimal risk	Patients who are not competent to give consent	Science Long termed	Very low	D

GRADE grading of recommendations assessment, development and evaluation
SIGN Scottish Intercollegiate Guidelines Network
Each of the five columns shows an order from a lower to a higher grade or from general to more specific items. However, the columns should not be seen as completely correlated to each other. Thus, e.g., in the risk column lines C and D are related only to children. "Chance" means the objective of the research intervention, i.e., the chance of a benefit for the individual research participant or at least for the group of patients with the same age or illness condition of the research patient, or the research results are expected to be applied currently (short termed) or are more of a basic nature with perhaps applicable findings in the future (long termed)

- Rid et al. (2010) produced after consultation with experts an empirically based scale of seven grades of injury and illustrated each grade with specific examples of the consequences of the injury and its duration and treatability (Rid et al. 2010) (see Table 58.3).
- Recently further proposals have been published in order to adapt control and monitoring of clinical research to the level of risk. Thus, a model of risk

Table 58.3 Magnitude of harms scale with illustrative examples[a] (From Rid et al. (2010) with permission of JAMA Copyright © (2010) American Medical Association. All rights reserved)

Examples and details of harms			
Examples of harms by magnitude	Effect/disability	Treatment	Duration
Negligible			
Mild nausea	Discomfort; can interfere with ability to pursue some minor life goals (e.g., eat)	May require medication	Minutes to several hours
Skin bruise or abrasion	Mild pain	Can require cleaning and coverage	Bruise or abrasion pain, minutes to several hours; healing, $\leq$10 days
Small			
Headache	Moderate pain, inability to pursue some minor (e.g., 1 day hiking) and some major (e.g., attend school) life goals	May require medication, rest, or both	Hours
Common cold	Discomfort, inability to pursue some minor (e.g., visit museum) and some major (e.g., work) life goals	May require medication, rest, or both	Several days
Moderate			
Uncomplicated bone fracture	Moderate pain, inability to pursue some minor life goals (e.g., play sports)	Requires some medication and wearing a cast	Fracture pain, hours; recovery, weeks to months
Moderate insomnia for 1 month	Annoying experience, inability to pursue some minor (e.g., meet friends) and some major (e.g., work) life goals	Can require lifestyle changes and medication	Weeks intermittently

Significant			
Ligament tear of knee with permanent instability	Moderate pain that interferes with pursuing some minor life goals (e.g., exercise); permanent instability precludes vigorous exercise and requires adaptation (e.g., seek new types of exercise)	Requires surgery and rehabilitation	Tear, hours to days; rehabilitation time following surgery, weeks to months
Intensive care for several weeks (assuming no sequelae)	Often intense pain and physical exhaustion, inability to perform activities of daily life and to pursue essentially all minor and major life goals		Weeks
Major			
Psychotic episode	Terrifying distortions of reality, changes in personality that undermine relationships, precludes performance of daily life activities and many minor and major life goals	Requires medication, can require adaptation of some major life goals (e.g., work)	Weeks to a month
Rheumatoid arthritis	Daily episodes of serious pain and permanent stiffness, unable to pursue some minor (e.g., vacation) and some major (e.g., work) life goals, sometimes unable to perform some activities of daily life	Requires aggressive medication, physiotherapy, requires major adaptation	Years
Loss of finger	Destabilizes hand, interferes with many activities of daily life, interferes with some minor and major life goals, requires adaptation, distressing transition period	None	Permanent

(*continued*)

Table 58.3 (continued)

Examples and details of harms			
Examples of harms by magnitude	Effect/disability	Treatment	Duration
Severe			
Major depression	Depressive episodes with hopelessness/worthlessness, loss of interest in usual activities, isomnia, and eating; can preclude performance of some daily life activities and some minor and major life goals; often baseline anxiety and low mood	Requires medication; requires adaptation of some major life goals (e.g., relationships)	Decades
Paraplegia	Inability to perform some activities of daily life, inability to pursue many minor (e.g., hiking) and some major (e.g., having children) life goals, often distressing transition period	Requires daily support and close clinical observation; requires major adaptation	Permanent
Catastrophic			
Severe dementia	Precludes performance of daily life activities and essentially all minor and major life goals, adaptation impossible, distressing transition period	Requires full-time care	Permanent
Death			

[a]Important factors that influence the magnitude of a harminclude associated experience (no sensory, nuissance, uncomfortable, distressing, suffering); burden of efforts to mitigate condition (low/moderate/high, weeks/months/permanent); inability to perform activities of daily life (partial/complete); inability to realize life goals (minor/major life goals, some goals in one category/some goals in both categories/all goals in one or both categories); duration (minutes/hours/day/weeks to months/years/permanent, intermitten/continuous); potential to adapt to new (residual) condition (minor/moderate/major adaptation, impossible to adapt); and burden of adaptation period (low/moderate/high). The examples were chosen based on input from 43 international experts in clinical research, research ethics, and risk assessment. The example have an illustrative function to show how the harm scale might be applied. Factors not mentioned in the description of an example are considered not relevant. It is assumed that the given harms occur in otherwise healthy, normal, average individuals (adults), which implies that the selected examples might fall into a different category on the harm scale in individuals who are not healthy, normal, or adults. No examples of economic or social harms are given due to their strong context dependence

estimation was introduced in 2010 into the discussion of the revision of the EU directive on clinical research that proposes to determine the two dimensions of severity and likelihood of impairments of health for defined groups of risks (sponsor and experience; product class; stage of development: pre- or post-marketing approval; scientific newness; characteristics of patients; method of trial) (Hartmann and Hartmann-Vareilles 2012). The German ADAMON project (Brosteanu et al. 2009) and the British community project(MRC/ and DH/ 2011) provide risk-adapted monitoring and control of clinical trials in three classes of risks with different intensity of control (and the bureaucratic effort).

To date, however, such proposals up to now are almost nothing more than constructs that can at most be useful for a rough structuring of the evaluation of the acceptability of benefits and risks of a research intervention and particularly of their relationship to each other. How much they can indeed achieve needs empirical validation.

Empirical Procedures for the Evaluation of the Benefit-Risk Relationship

- In contrast to the abovementioned proposals, such an empirical study of a more differentiated three-stage taxonomy was done ex post from all research applications of the year 2006 at the university hospital in Lübeck (Hüppe and Raspe 2011):

> "At the first step of analysis the identified consequences were differentiated according to their effects (positive or negative) in chances of benefit and risks of injury (with regard to addressee, relation to the study, relevance, extent, likelihood of occurrence, start of occurrence, sustainability, level of evidence)." "The second step of the analysis outlines the chances of benefit and risks of damages with regard to the respective addressee or concerned person into three partial quantities": potential self- benefit or self-injury; group benefit or group injury; benefit for others or injury for others. In step 3 of the analysis the detailed assessment of the features of individual chances of benefit and risks of injury" with regard to the criteria named under step 1 was carried out. Subsequently it was stated by two steps of balancing "whether the entire balance yields a net benefit"." In order to reach a concluding positive point balance a net benefit must be found." Such a "proceduralisation of the analysis of chances and risks can increase the transparency of the performed processes of analysis and comparison. The communication among researchers and ethics committees and among the members of the ethics committee will be facilitated in controversial research projects, the standardization and harmonisation of the consulting processes of the ethics committees will be supported.

- Recently the estimated risks of two research interventions, an allergy test and a liver biopsy, were compared with empirically assessed risks of everyday activities. In the first example, the magnitude and the likelihood of occurrence of the risks of allergy testing are below those of comparable everyday activities (e.g., sports or work), i.e., seem to be acceptable (Fig. 58.1).

However, the liver biopsy as a research intervention is ethically questionable because some of its risks are above comparable everyday activities (Fig. 58.2).

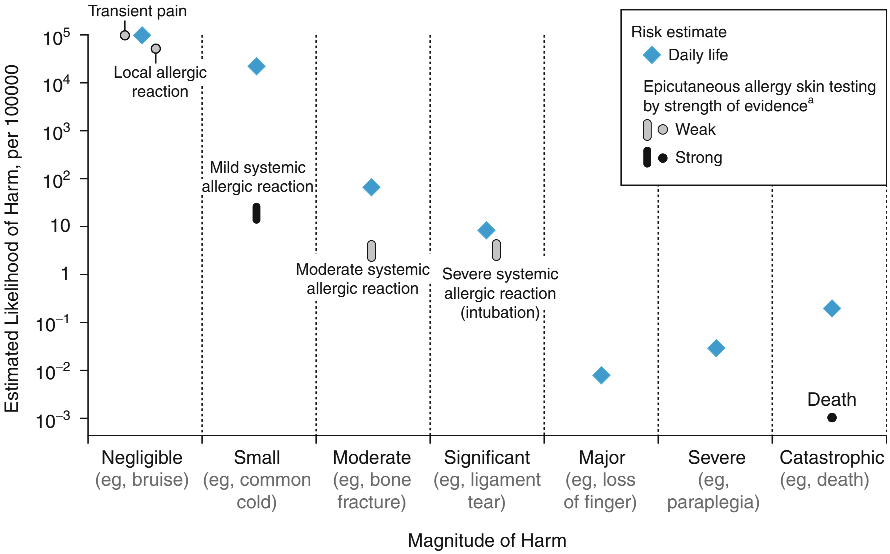

Fig. 58.1 Estimated risks of epicutaneous allergy skin testing (per 100,000): transient pain (negligible), approximately 100,000; local allergic reaction (negligible), approximately 50,000; mild systemic allergic reaction (small), 11–30; moderate or severe systemic allergic reaction (moderate or significant), 2–5; and death (catastrophic), approximately 0 (1 case report). Daily life risks in the United States (per 100,000): bruise (negligible), approximately 100,000 (all age groups); common cold (1 day [small]), approximately 22,000 (children); bone fracture or dislocation (surfing contest [moderate]), approximately 70 (adults); complete ligament tear of knee (sports practice [significant]), approximately 8 (adolescents); loss of 1 finger (workday in service sector [major]), approximately 0.008 (adults); paraplegia (day of skiing [severe]), approximately 0.03 (all age groups); and death (riskier car trip [catastrophic]), approximately 0.2 (adolescents/adults).[a]Span of elongated *data markers* indicates range of estimated risk (From Rid et al. (2010) with permission of JAMA Copyright © (2010) American Medical Association. All rights reserved)

This is a considerable step towards an empirical foundation of risk assessment. However, it is an open question whether it will gain acceptance because it is a very elaborate procedure.

Contextual Influences

An acceptable or reasonable benefit-risk relationship is understood usually as a justified relationship between benefits and risks. At the same time, it should be considered that the estimation of a benefit-risk relationship as justified also depends upon normative values.

Example

In studies with potential individual benefit but more than minimal risks, e.g., trials of new methods of vaccination, it must be decided whether the benefit-risk relationship is ethically acceptable in patients not competent to give

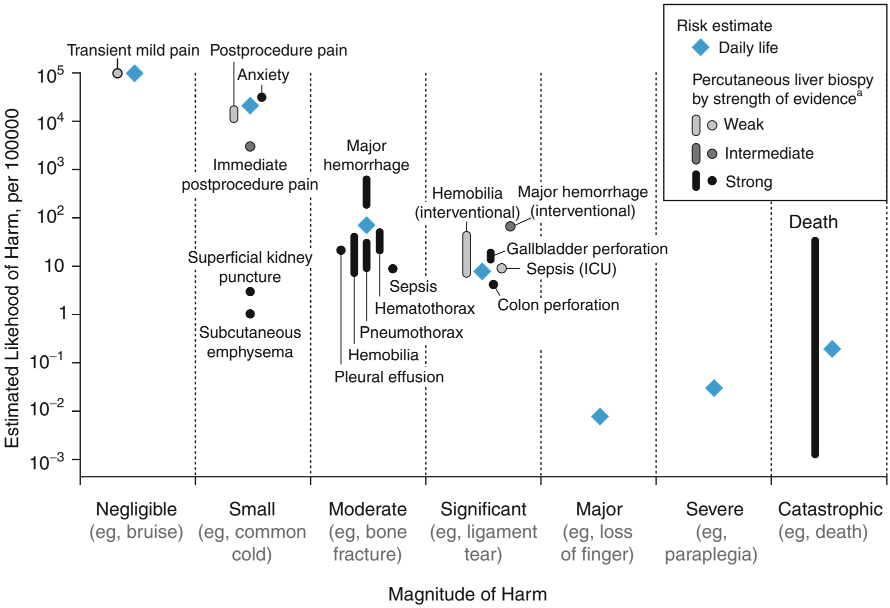

Fig. 58.2 Estimated risks of percutaneous liver biopsy (per 100,000): anxiety (small), 31,000; transient mild pain (negligible), approximately 100,000; immediate postprocedure pain (small), approximately 3,000; postprocedure pain (small), approximately 10,000–20,000; superficial kidney puncture (small), 3 subcutaneous emphysema (small), 1 pleural effusion (moderate), 21 hematothorax (moderate), 18–63 pneumothorax (moderate), 8–35 major hemorrhage requiring transfusion (moderate), 160–733 hemobilia requiring conservative treatment (moderate), 6–50 sepsis requiring antibiotic treatment (moderate), 9 major hemorrhage requiring interventional radiography/surgery (significant), 67 hemobilia requiring interventional treatment (significant), 6–50 sepsis requiring intensive care unit (ICU) treatment (significant), 9 gallbladder perforation (significant), 12–22 colon perforation (significant), 4 death (catastrophic), 0–40. For daily life risks see Figure 58.1 legend.[a]Span of elongated *data markers* indicates range of estimated risk (From Rid et al. (2010) with permission of JAMA Copyright © (2010) American Medical Association. All rights reserved)

informed consent and with currently untreatable disease conditions such as advanced stages of Alzheimer disease (analogous to the reasoning for experimental treatment in patients with final stages of cancer). Some family members might view the risk of burdens for the patient as considerable and are not convinced of the chance of a remission and therefore hope for a peaceful end for the patient; other relatives – in harmony with a possible advance directive or presumptive will of the patient – or also members of the ethics committee might estimate the potential benefit of an attenuation of symptoms or a prolongation of life to be much greater than the potential burdens.

Thus, for now it remains largely a personal evaluation of the participants judging the ethical acceptability of a research intervention. This evaluation is filled with uncertainties and is thereby open to contextual influences. But "what cannot be objectified can at least be controlled by procedures: one attempts to distribute difficulties onto several shoulders and hopes that the collectivity of judgments may lead to a result that is more acceptable. However, the basic difficulties of a balance between benefits and risks are not solved by this but are only regulated by procedure" (Wiesing 2011).

In order to control and to minimize these contextual influences, a three-step procedure will be formulated here:

1. The *researcher* must give his/her reasons for considering that the relationship of potential risks and burdens to the expected benefit of a planned research project is acceptable, i.e., that it is reasonable and justified.
2. The appropriate *ethics committee* must investigate this relationship and the evaluation of the researcher with regard to ethical and legal norms, if necessary with additional professional expertise. It should communicate its reasons – at least in research with patients not competent to give informed consent – not only in cases of rejection but also in cases of approval of the research plan and particularly with regard to the ethical argumentation of the researcher.
3. Finally the *potential research participant* or his/her authorized substitute must evaluate the reasons of the institutionally acceptable established benefit-risk relationship of the planned research intervention with regard to his/her own idiosyncrasies, values, and interests; after that he/she can consent if he/she feels that the benefit-risk relationship seems to be acceptable for himself/herself.

Conclusion

The evaluation of the benefit-risk relationship of a medical research intervention is only probabilistically possible and is open for contextual influences, because the criteria of benefits and of risks are often only insufficiently quantitatively defined. Whereas in a medical standard intervention the benefit-risk estimation is focused almost only on the individual, in a research intervention the societal benefit-risk evaluation must be added. The question remains whether it is at all possible and, if so, how individual benefits and risks can be balanced against societal benefits and risks, particularly if the research intervention contains more than minimal risks.

Only in the past decade researchers have begun to investigate determinants of the benefit-risk evaluation in ethics committees, to develop evaluation procedures (Curtin and Schulz 2011), and to systematize in a framework (Rid and Wendler 2011). Algorithmic attempts to structure the evaluation process should standardize the evaluation. However, for now only a pragmatic solution will be possible that will validate the result in three steps (researcher, ethics committee, potential research participant).

Cross-References

▶ Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
▶ Strengthening Self-Determination of Persons with Mental Illness

References

Brosteanu, O., Houben, P., Ihrig, K., Ohmann, C., Paulus, U., Pfistner, B., Schwarz, G., Strenge-Hesse, A., & Zettelmeyer, U. (2009). Risk analysis and risk adapted on-site monitoring in noncommercial clinical trials. *Clinical Trials, 6*, 585–596.

Connell, C. M., Shaw, B., Holmes, S. B., & Forster, N. L. (2001). Caregivers' attitudes toward their family members' participation in Alzheimer disease research: Implications for recruitment and retention. *Alzheimer Disease and Associated Disorders, 15*, 137–145.

Curtin, F., & Schulz, P. (2011). Assessing the benefit: Risk ratio of a drug–randomized and naturalistic evidence. *Dialogues in Clinical Neuroscience, 13*, 183–190.

Deutscher Bundestag. (2004). Arzneimittelgesetz (AMG-04) (1976/2004) incl. 12. Novelle.

Deutscher Bundestag. (2010). Sozialgesetzbuch (SGB), Fünftes Buch (V), Gesetzliche Krankenversicherung. Zuletzt geändert durch Art. 2 G v. 22.12.2010 I 2309. http://wwwzahnaerzte-wlde/webtest/Internet_ZAEKnsf/66a4836bca0a3292c125741f00363d72/2ea079ee49c03956c12578a800451723/$FILE/SGB%20V%20Stand%2003-2011pdf

Emanuel, E., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? *JAMA: The Journal of the American Medical Association, 283*, 2701–2711.

Europarat. (1997). Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine' (No. 164) (CHRB-97)

Europarat. (2005). Additional protocol to the convention on human rights and biomedicine, concerning Biomedical Research (No. 195) (AD-05) http://conventionscoeint/Treaty/en/Treaties/Html/195htm

Europarat. (2011). Explanatory report – Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine. http://conventionscoeint/treaty/en/Reports/Html/164htm

Fisher, C. B., Kornetsky, S. Z., & Prentice, E. D. (2007). Determining risk in pediatric research with no prospect of direct benefit: Time for a national consensus on the interpretation of federal regulations. *The American Journal of Bioethics, 7*, 5–10.

Friedman, A., Robbins, E., Wendler, D. (2010). Which benefits of research participation count as 'Direct'? *Bioethics*. PMID: 20497168 [PubMed – as supplied by publisher] PMCID: PMC2945615. Accessed 17 Nov 2011.

Gefenas, E. (2007). Balancing ethical principles in emergency medicine research. *Science and Engineering Ethics, 13*, 281–288.

Gigerenzer, G. (2006). Einfache Heuristiken für komplexe Entscheidungen. In: Präsident der Berlin-Brandenburgischen Akademie der Wissenschaften (Ed.) *Mathematisierung der Natur: Streitgespräche in den Wissenschaftlichen Sitzungen der Versammlung der Berlin-Brandenburgischen* (pp. 37–44). Berlin: Akademie der Wissenschaften

Hartmann, M., & Hartmann-Vareilles, F. (2012). Concepts for the Risk-Based Regulation of Clinical Research on Medicines and Medical Devices. *Drug Information Journal, 46*, 545–554.

Heinrichs, B. (2007). *Forschung am Menschen. Elemente einer ethischen Theorie biomedizinischer Humanexperimente*. Berlin/New York: Walter de Gruyter.

Helmchen, H. (2002). Biomedizinische Forschung mit einwilligungsunfähigen Erwachsenen. In J. Taupitz (Ed.), *Das Menschenrechtsübereinkommen zur Biomedizin des Europarates – Taugliches Vorbild für eine weltweit geltende Regelung?* (pp. 83–115). Berlin/Heidelberg/New York: Springer.

Helmchen, H., & Lauter, H. (1995). Dürfen Ärzte mit Demenzkranken forschen? Analyse des Problemfeldes Forschungsbedarf und Einwilligungsproblematik. New York/Stuttgart: Thieme.
Hüppe, A., & Raspe, H. (2011). Mehr Nutzen als Schaden? Nutzen- und Schadenpotenziale von Forschungsprojekten einer Medizinischen Fakultät – eine empirische Analyse. *Ethik in der Medizin, 23*, 107–121.
Magnus, D., & Merkel, R. (2007). Normativ-rechtliche Grundlagen der Forschung an Nichteinwilligungsfähigen. In J. Boos, R. Merkel, H. Raspe, & B. Schöne-Seifert (Eds.), *Nutzen und Schaden aus klinischer Forschung am Menschen. Abwägung, Equipoise und normative Grundlagen*. Köln: Deutscher Ärzteverlag.
Maier, W., Wagner, M., & Stingelin, N. (2013). Ethische Aspekte der molekulargenetischen Forschung. In H. Helmchen (Ed.), *Ethik psychiatrischer Forschung*. Heidelberg: Springer.
Mastwyk, M., Ritchie, C. W., LoGiudice, D., Sullivan, K. A., & Macfarlane, S. (2002). Carers' impressions of participation in Alzheimer's disease clinical trials: What are their hopes? And is it worth it? *International Psychogeriatrics, 14*, 39–45.
MRC/, DH/, MHRA. (2011). Joint project. Risk-adapted approaches to the management of clinical trials of investigational medicinal products. In: http://www.mhra.gov.uk/home/groups/l-ctu/documents/websiteresources/con111784.pdf (Ed.). Accessed 8 July 2013.
Rajczi, A. (2004). Making risk-benefit assessments of medical research protocols. *The Journal of Law, Medicine & Ethics, 32*, 338–348.
Raspe, H. (2012). persönliche Mitteilung: Explanatorisches vs pragmatisches Wissen. In email vom 26 June 2012 (Ed).
Resnik, D. B. (2005). Eliminating the daily risks standard from the definition of minimal risk. *Journal of Medical Ethics, 31*, 35–38.
Rid, A., & Wendler, D. (2011). A framework for risk-benefit evaluations in biomedical research. *Kennedy Institute of Ethics Journal, 21*, 141–179.
Rid, A., Emanuel, E., & Wendler, D. (2010). Evaluating the risks of clinical research. *JAMA: The Journal of the American Medical Association, 304*, 1472–1479.
Rosenbaum, L. (2012). How much would you give to save a dying bird? Patient advocacy and biomedical research. *The New England Journal of Medicine, 367*, 1755–1759.
Shah, S., Whittle, A., Wilfond, B., Gensler, G., & Wendler, D. (2004). How do institutional review boards apply the federal risk and benefit standards for pediatric research? *JAMA: The Journal of the American Medical Association, 291*, 476–482.
Simonsen, S. (2009). Acceptable risk and the requirement of proportionality in European Biomedical Research Law. What does the requirement that biomedical research shall not involve risks and burdens disproportionate to its potential benefits mean? In Trondheim: Norwegian University of Science and Technology (NTNU).
Simonsen, S. (2012). *Acceptable risk in biomedical research. European perspectives*. Dordrecht/Heidelberg/London/New York: Springer.
Sofaer, N., Jafarey, A., Lei, R. P., Zhang, X., & Wikler, D. (2007). Unconditional compensation: Reducing the costs of disagreement about compensation for research subjects. *Eastern Mediterranean Health Journal, 13*, 6–16.
Terwey, J. H. (2007). *Die Struktur ethisch relevanter Kategorien medizinischer Forschung am Menschen*. Med Dissertation.
Vollmann, J. (2000). "Therapeutische" versus "nicht-therapeutische" Forschung – eine medizinethische plausible Differenzierung? *Ethik in der Medizin, 12*, 65–74.
Welie, S. P. K., & Berghmans, R. L. P. (2006). Inclusion of patients with severe mental illness in clinical trials: Issues and recommendations surrounding informed consent. *CNS Drugs, 20*, 67–83.
Wendler, D. (2008). Is it Possible to Protect Pediatric Research Subjects without Blocking Appropriate Research? *Journal of Pediatrics, 152*, 467–470.
Wendler, D. (2009). Minimal risk in pediatric research as a function of age. *Archives of Pediatrics & Adolescent Medicine, 163*, 115–118.
Wendler, D., & Emanuel, E. J. (2005). What is a "minor" increase over minimal risk? *Journal of Pediatrics, 147*, 575–578.

Wendler, D., & Miller, F. G. (2007). Assessing research risks systematically: The net risks test. *Journal of Medical Ethics, 33*, 481–486.

Wendler, D., Belsky, L., Thompson, K. M., & Emanuel, E. J. (2005). Quantifying the federal minimal risk standard: Implications for pediatric research without a prospect of direct benefit. *JAMA: The Journal of the American Medical Association, 294*, 826–832.

Wendler, D., Krohmal, B., Emanuel, E. J., Grady, C., & Group, E. (2008). Why patients continue to participate in clinical research. *Arch Intern Med, 168*, 1294–1299.

Wiesing, U. (2011). Comments on "Ethics of clinical resarch with mentally ill persons". Letter to the author

World Medical Association. (2008). Declaration of Helsinki (1964/2008). http://wwwwmanet/en/30publications/10policies/b3/17cpdf

Zentrale Ethikkommission bei der Bundesärztekammer. (1997). Stellungnahme "Zum Schutz nicht-einwilligungsfähiger Personen in der medizinischen Forschung". *Deutsches Ärzteblatt, 94*, B811–B812.

Section XIII

Ethics in Neurosurgery

Ethics in Neurosurgery

59

Marcos Tatagiba, Odile Nogueira Ugarte, and Marcus André Acioly

Contents

Abstract

In this section, we provide a comprehensive overview of the Ethics in Neurosurgery from the basic concepts of morality and ethical theory to the emergence of neuroethics. We further discuss development of new ethical dilemmas as technological advances come into clinical use and how general and medical society should be gathered in that ethical discussion.

The concept of morality addresses what is and is not socially acceptable. Morality deals with social conventions in terms of what is considered right or wrong, based on an implicit consensus among members of a specific community. Ethical theory, in turn, refers to an analysis of morality, the study and understanding of its nature, as well as its function (Beauchamp and Childress 1994). Medical ethics refers to the

M. Tatagiba (✉)
Department of Neurosurgery, Eberhard-Karls University Hospital, Tübingen, Germany
e-mail: marcos.tatagiba@med.uni-tuebingen.de

O.N. Ugarte
Division of Neurosurgery, Andaraí Federal Hospital, Rio de Janeiro, Brazil

Postgraduation Program in Neurology, Federal University of the State of Rio de Janeiro, Rio de Janeiro, Brazil

M.A. Acioly
Division of Neurosurgery, Fluminense Federal University, Niterói, Rio de Janeiro, Brazil

Division of Neurosurgery, Andaraí Federal Hospital, Rio de Janeiro, Brazil

Department of Neurosurgery, Eberhard–Karls University Hospital, Tübingen, Germany

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_61,

study of morality as applied to medicine. The analysis of moral issues in terms of their relationship to professional behavior has led to the creation of professional ethics codes, among them the code of medical ethics, in order to guide and to create norms of conduct for medical practice. In this context, recent advances in neuroscience have sparked the creation of a new subdivision of this field, neuroethics.

The terms proposed by the codes of professional ethics may vary from country to country, but, in general, they support principles of autonomy, of beneficence, of non-maleficence, and of justice. Of the four principles, the principle of autonomy is perhaps the one most discussed in the literature about ethics, especially in the literature about the doctor-patient relationship and obtaining informed consent (Faden and Beauchamp 1986). Respecting patients' autonomy is one of the pillars of modern bioethics in the West. Autonomy recognizes the people's right to make decisions about their own treatments, free from any pressure or manipulation (Beauchamp and Childress 1994).

This attitude towards the principle of autonomy stands in opposition to the classic paternalist model, in which the doctor alone makes all the decisions. Through the principle of autonomy, patients have the right to receive clear information about their treatment and to make decisions based on their own beliefs and values. The doctor should help the patient to make choices and avoid coercion and persuasion, beyond providing relevant information in a comprehensible way, encouraging questions, and clarifying the management options that are available (Etchells et al. 1996b). Thus, autonomy can only be fully exercised when doctors fulfill their role to provide information. Establishing a good doctor-patient relationship is fundamental to any therapy's success, in the sense that the doctor should know and respect patient's expectations. The professional's goal should be one of kindness, considering that patient's interests and well-being come first. However, patients' expectations can vary, and the way in which any particular doctor deals with the questions involved in the therapeutic process may be very individualized.

There are many studies about patient preferences in the decision-making process. The results vary and are correlated with cultural, demographic, religious, and socioeconomic characteristics (Delgado et al. 2010; Coulter and Jenkinson 2005; Levinson et al. 2005; Robinson et al. 2001; Bruera et al. 2001; Pang 1999; Blackhall et al. 1995). Not all patients want to take control of the issues involved in their treatment. Some prefer a participatory model of decision-making and ask for the doctor's involvement, while others transfer their decisions to someone they trust (Delgado et al. 2010; Coulter and Jenkinson 2005; Levinson et al. 2005). In other words, if patients want to, they can delegate responsibility to a member of their family or ask for advice from the doctor, without this constituting a violation of their autonomy (Etchells et al. 1996; Lazar et al. 1996). Preferences also vary when research is performed on healthy or sick people and may depend on the nature of the illness itself (Delgado et al. 2010; Ruhnke et al. 2000).

Informed consent is an instrument that, when applied properly, guarantees the full exercise of autonomy. Originally, the term was referred to obtaining the patient's consent to perform a procedure. This consent might be verbal or via a form signed by the patient called a consent form, which these days fulfills

a purpose additional to its ethical one, having acquired legal value, as an instrument of protection. Doctors, however, still have an obligation to inform, and the patient's right to receive clear information should not be viewed merely as protection against potential litigation.

Some publications in the area of neurosurgery have considered patients' increased participation in the way the treatment itself is conducted as a violation of the traditional doctor-patient relationship. Moreover, the legal focus given to the application of terms of consent can cause neurosurgeons to "hyper-inform" their patients, as a "matter of safety" (Schmitz and Reinacher 2006). Doctors need to develop a good relationship with their patients so they can know and understand their expectations. This can be achieved with an honest conversation, in which patients are encouraged to resolve specific questions, but also by presenting a consent form giving the opportunity for the doctor to get to know patients better and the sociocultural context into which they are inserted.

Because of the need to guide neurosurgeons through contemporary ethical dilemmas, the World Federation of Neurosurgical Societies developed a statement of ethics in neurosurgery (Umansky et al. 2011). The document presents a set of directives about good practice in neurosurgery, with a special focus on maintaining a good relationship with patients and their families. It emphasizes that the surgeon's duty is to listen to patients, to respect their choices, and to clarify any doubts. The neurosurgeon has the responsibility to clearly inform and to obtain patients' consent before performing surgical procedures. The importance of good communication with family members, of respecting cultural differences, and of the need to act with kindness is all themes that the document broaches. In addition, the text underscores the relevance of good clinical practice and the need for doctors to continually update their knowledge and studies, as well as exercising care in adopting new technologies that have not yet been scientifically proven. The statement of ethics includes a discussion of what constitutes excellence in clinical practice and reminds us that, at a time when technology is constantly changing, human relations cannot be neglected.

Recent progress in neuroscience and the increased understanding of the nervous system's physiology have led to new ethical dilemmas. The brain has come to be understood as the organ in which the essence of our individuality resides. We increasingly understand our personality, our feelings, our behavior, our memories, and our perceptions as derivatives of brain function. Research in neuroscience challenges the very concept of what human nature is (Illes and Bird 2006). In this context, neuroscience offers a unique possibility to access the human brain. While basic advances in cognitive neuroscience and brain imaging have promoted a potential application in clinical practice and in neurosurgical research, the access to the brain that is made possible by neurosurgeons offers a chance to achieve significant advances in cognitive neuroscience (Tatagiba et al. 2007).

In practical terms, we can use the example of the awake surgery for resecting brain tumors. The distinctive feature of this type of surgery lies in the ability to improve tumor resection, preserving adjacent areas involved in cognitive function. Since various cognitive functions, such as language, memory and orientation, and

perception, can be tested during the operation, anthropological questions arise. What functions are really essential to being "human"? What functions make up the patient's individual personality? Making a hierarchy of these cognitive functions prior to the operation may be necessary in order to explicitly determine what the operation ought to preserve. This process conjures up other questions; are there cognitive functions that should not be damaged in some way during surgery? Or how does one make a decision when tumor resection might affect memory or language? Or, again, is an almost total tumor resection associated with a discrete reduction in memory preferable to a total resection with more significant cognitive damage (Clausen and Gharabaghi 2007)?

Neuroethics has arisen from the need to study the ethical, legal, and social implications of neuroscientific advances. The existence of functional exams allowing for the identification and location of complex thought processes, devices for deep brain stimulation, medical treatments and operations for psychiatric illnesses, and treatments for degenerative diseases with stem cell transplants are some of the challenges that neuroethics faces today (Lombera and Illes 2009). Thus, neuroethics is not simply a branch of bioethics, but a new science that deals with the definition of what it means to be human, with free will and self-knowledge (Levy 2011).

Therefore, the discipline of neuroethics can be subdivided into two branches: the ethics of neuroscience and the neuroscience of ethics. The *ethics of neuroscience* is the application of ethical principles to the studies and technologies developed by neuroscientists. The *neuroscience of ethics* involves what studying the mind tells us about the nature of morality (Roskies 2002).

One of the new ethical challenges prompted by scientific progress is the possibility for cognitive control through drugs or brain implants. The use of memory-modifying technologies seems a promising way to treat victims of post-traumatic stress. Using beta-blockers has allowed us to edit our memories. Since our past is a part of how we construct our identity, to what degree is the use of these new technologies socially acceptable (Erler 2011)? Methods to pharmacologically edit memory may be justifiable in the case of survivors of traumatic events, but using them in indiscriminate or "cosmetic" ways is discouraged (Erler 2011). We can enhance memory and attention pharmacologically by using modulators of AMPA glutamate receptors (*α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid*). Using neural implants to modulate brain function is another facet of external control over cognitive function. These new instruments bring with them moral dilemmas that demand a broad debate from the scientific community and broader society (Illes and Bird 2006).

The attempt to obtain surgical control of the complex mechanisms involved in cognition and the nature of our thought processes is not new in the history of medicine. Psychosurgery began in the early twentieth century, with Egas Moniz's development of the frontal lobotomy and Freeman's popularization of it (Hollingham 2009). The initial enthusiasm for the technique and the media attention paid to it led to an excess number of surgical indications, which only stopped after scientific publications about its harmful effects began to appear.

This historical example shows the importance of having a philosophical debate about the moral and ethical implications of every technological advance before popularizing it (Gilbert and Ovadia 2011). At the moment, the media has been heaping praise on deep brain stimulation devices, whether to treat psychiatric disorders or to control movement disorders, creating a natural demand for these new technologies, which society views with extreme optimism. Deep brain stimulation is, nevertheless, an invasive procedure that can lead to side effects in terms of cognitive behavior – and one whose mechanism is not entirely clear. It should only be indicated after a careful analysis of the moral questions involved (Gilbert and Ovadia 2011).

On the whole, neuroscience is a constantly changing field. As medical knowledge progresses, it creates new concepts and modifies clinical practice. In today's world, hospitals are incorporating new technologies on an almost daily basis, more quickly than ever before. The result of this state of constant change is that doctors constantly need to learn new techniques. As they are incorporated into the profession, new technological advances bring new ethical dilemmas. Society and the scientific world must maintain a continuous dialogue about the moral implications of these new technologies. Adopting health policies that meet the basic principles of bioethics depends on this debate.

References

Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. New York: Oxford University Press.

Blackhall, L. J., Murphy, S. T., Frank, G., Michel, V., & Azen, S. (1995). Ethnicity and attitudes toward patient autonomy. *Journal of American Medical Association, 274*(10), 820–825.

Bruera, E., Catherine, S., Calder, K., Palmer, L., & Benisch-Tolley, S. (2001). Patient preferences versus physician perceptions of treatment decisions in cancer care. *Journal of Medical Ethics, 19*(11), 2883–2885.

Clausen, J., & Gharabaghi, A. (2007). Neuroethics and awake surgery: anthropological and ethical implications of interactive brain surgery. In Tatagiba, M., Pavlova, M., Gharabaghi, A., & Sokolov, A. A. (Eds.), *Combining neuroscience with neurosurgery: proceedings of the 1st international symposium on cognitive neurosurgery* (pp. 187–188). Kirchentellinsfurt: Knirsch Verlag.

Coulter, A., & Jenkinson, C. (2005). European patient's views on the responsiveness of health systems and healthcare providers. *European Journal of Public Health, 15*(4), 355–360.

Delgado, A., López-Fernández, L. A., Luna, J. D., Cuesta, L. S., Garrido, N. G., & González, A. P. (2010). Patients' expectations about decision-making in terms of different health issues. [Expectativas de los pacientes sobre la toma de decisiones ante diferentes problemas de salud]. *Gaceta Sanitaria, 24*(1), 66–71 (Spanish)

Erler, A. (2011). Does memory modification threaten our authenticity? *Neuroethics, 4*, 235–249.

Etchells, E., Sharpe, G., Dykeman, M. J., Meslin, E. M., & Singer, P. A. (1996a). Bioethics for clinicians: 4 voluntariness. *Canadian Medical Association Journal, 155*(8), 1083–1086.

Etchells, E., Sharpe, G., Walsh, P., Williams, J. R., & Singer, P. A. (1996b). Bioethics for clinicians: 1 consent. *Canadian Medical Association Journal, 155*(2), 177–180.

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. New York: Oxford University Press.

Gilbert, F., & Ovadia, D. (2011). Deep brain stimulation in the media: over-optimistic portrayals call for a new strategy involving journalists and scientists in ethical debates. *Frontiers in Integrative Neuroscience, 5*, 16.
Hollingham, R. (2009). *Blood and guts: A history of surgery*. New York: Thomas Dunne Books.
Illes, J., & Bird, S. J. (2006). Neuroethics: Modern context for ethics in neuroscience. *Trends in Neurosciences, 29*(9), 511–517.
Lazar, N. M., Greiner, G. G., Robertson, G., & Singer, P. (1996). Bioethics for clinicians: 5 substitute decision-making. *Canadian Medical Association Journal, 155*(10), 1435–1437.
Levinson, W., Kao, A., Kuby, A., & Thisted, R. A. (2005). Not all patients want to participate in decision making. A national study of public preferences. *Journal of General Internal Medicine, 20*(6), 31–535.
Levy, N. (2011). Neuroethics: A new way of doing ethics. *American Journal of Bioethics Neuroscience, 2*(2), 3–9.
Lombera, S., & Illes, J. (2009). The international dimensions of neuroethics. *Developing World Bioethics, 9*(2), 57–64.
Pang, S. M. (1999). Protective truthfulness: The Chinese way of safeguarding patients in informed treatment decisions. *Journal of Medical Ethics, 25*, 247–253.
Robinson, A., & Thomson, R. (2001). Variability in patient preferences for participating in medical decision making: Implication for the use of decision support tools. *Quality in Health Care, 10*(suppl I), 34–38.
Roskies, A. (2002). Neuroethics for the new millennium. *Neuron, 35*, 21–23.
Ruhnke, G. W., Wilson, S. R., Akamatsu, T., Kinoue, T., Takashima, Y., Goldstein, M. K., Koenig, B. A., Hornberger, J. C., & Raffin, T. A. (2000). Ethical decision making and patient autonomy. *Chest, 118*(4), 1172–1182.
Schmitz, D., & Reinacher, P. C. (2006). Informed consent in neurosurgery – translating ethical theory into action. *Journal of Medical Ethics, 32*, 497–498.
Tatagiba, M., Pavlova, M., Gharabaghi, A., & Sokolov, A. A. (Eds.). (2007). *Combining neuroscience with neurosurgery: proceedings of the 1st international symposium on cognitive neurosurgery* (p. 3). Kirchentellinsfurt: Knirsch Verlag
Umansky, F., Black, P. L., DiRocco, C., Ferrer, E., Goel, A., Malik, G. M., Mathiesen, T., Mendez, I., Palmer, J. D., Juanotena, J. R., Fraifeld, S., & Rosenfeld, J. V. (2011). Statement of ethics in neurosurgery of the world federation of neurosurgical societies. *World Neurosurgery, 76*(3–4), 239–247.

Neurosurgery: Past, Present, and Future 60

Marcos Tatagiba, Odile Nogueira Ugarte, and Marcus André Acioly

Contents

Abstract

The practice of neurosurgery as we know it today is the result of centuries of evolution. Progress does not occur in a linear and methodical way, but rather in many cases stems from repeated episodes of trial and error. Although intellectual curiosity and the investigative spirit have been present throughout history, years of darkness and stagnation often follow fruitful periods of innovation. In this chapter, we provide an overview of the history of neurosurgery, from trepanations in the Neolithic period to the modern integrated operating rooms of today, with some considerations of the future prospects of the field.

M. Tatagiba (✉)
Department of Neurosurgery, Eberhard-Karls University Hospital, Tübingen, Germany
e-mail: marcos.tatagiba@med.uni-tuebingen.de

O.N. Ugarte
Division of Neurosurgery, Andaraí Federal Hospital, Rio de Janeiro, Brazil

Postgraduation Program in Neurology, Federal University of the State of Rio de Janeiro, Rio de Janeiro, Brazil

M.A. Acioly
Division of Neurosurgery, Fluminense Federal University, Niterói, Rio de Janeiro, Brazil

Division of Neurosurgery, Andaraí Federal Hospital, Rio de Janeiro, Brazil

Department of Neurosurgery, Eberhard–Karls University Hospital, Tübingen, Germany

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_62,

The Past

The oldest neurosurgical practice we know of is trepanation, a term derived from the Greek word *trypanon* (borer), which involves making openings in the cranium by removing bone fragments. Trepanned crania have been found on every continent, with the oldest known samples being found in the Ukraine dating to the Mesolithic period (Carod-Artal and Vázquez-Cabrera 2004). The fact that trepanation was commonly practiced by civilizations across the globe says a lot about human nature. Different societies that were not in communication with one another, as far as we can tell, nevertheless developed similar cranium perforation techniques. But what was the goal of these primeval surgeries? We believe that people performed trepanations for a variety of motives, such as for ritualistic purposes or for religious or therapeutic reasons. Instances of postmortem trepanations that enabled people to use bone fragments as amulets have been found in France (Liu et al. 2003a).

In South America, a vast collection of trepanned crania that were preserved by the arid climate of the Peruvian coast have been uncovered. Approximately 5–6 % of the 10,000 mummies found in Peru show evidence of trepanations, including many which exhibit signs of bone remodeling, indicating that the person survived the procedure. Archeological findings come from the Paracas, Nazca, Chimú, and Huari cultures as well as from the Incan empire (Carod-Artal and Vázquez-Cabrera 2004). These primitive surgeons apparently had some empirical knowledge and had learned to avoid perforating dura mater sinuses and areas covered with muscles, thereby avoiding hemorrhage (Andrushko and Verano 2008). Pre-Columbian cultures developed a variety of trepanation techniques, including circular incisions, scraping, transverse cuts, and perforations, that were executed with the help of chisels made of obsidian, silex, and *tumis*. The *tumi* was a T-shaped instrument with a semicircular blade that is the symbol of Peruvian medicine today. Most Peruvian trepanations were located on the left side, in the temporoparietal area, probably in order to treat lesions caused by hand-to-hand combat with a skilled right-handed opponent (Andrushko and Verano 2008; Liu et al. 2003a).

Surprisingly, a group of Peruvian neuroscientists performed two surgical procedures in the 1940s and 1950s using sterilized pre-Columbian instruments. Whatever the moral and ethical implications of these acts, they did prove that it was possible to perform trepanation using primitive instruments (Carod-Artal and Vázquez-Cabrera 2004; Marino and Gonzalles-Portillo 2000).

The oldest known medical text to contain references to the practice of neurosurgery is the Edwin Smith Papyrus, given this name because Edwin Smith, an American living in Egypt, acquired it in 1862. The papyrus text dates from 1700 B. C., but it is actually a copy of an older text, from around 3000–2500 B.C., by an unknown author. The papyrus contains 48 surgical cases, each of which includes a description of the exam, diagnosis, and treatment. Of these, 27 were cranial lesion cases (Feldman and Goodrich 1999).

Hippocrates (460–377 B.C.), whose oath is mechanically repeated by students today, is considered the father of modern medicine. Hippocrates developed the

concept of careful observation applied to clinical practice and emphasized the importance of a detailed anamnesis. Based on his observations of numerous cranial traumas, most stemming from battles, Hippocrates wrote his *De Capitis Vulneribus*, a treatise about the diagnosis and treatment of cranial lesions. His system for classifying fractures covered five subtypes, with the treatment indicated being based on lesion type rather than the patient's symptoms. This practice persisted for a long time until the development of the concept of cerebral localization. Hippocrates argued against trepanations for depressed skulls and comminuted fractures, probably because the objective of trepanation was to create a drainage opening for blood or fluid. Hence, since the fracture had already created such an orifice, these cases would not have required trepanations. Hippocrates saw suppuration as a natural and desirable process for wound healing; this concept remained unaltered for many years (Panourias et al. 2005).

Because dissection was taboo at that time, Hippocrates's observations came from clinical practice. Herophilus (335–280 B.C.) performed the first methodical and regular dissections of the human body as an object of anatomical study. He not only dissected cadavers, but was also adept at vivisecting criminals. Herophilus cruelly contravened the limits of morality and humanity, but he made an unparalleled contribution to anatomical studies. Herophilus provided a detailed description of the dura mater sinuses (*torcular Herophili*) of the ventricular system and of the choroid plexus. He contradicted Aristotelian teaching by proclaiming that the brain was the home of the soul. He differentiated tendons from nerves and motor nerves from sensory nerves. After Herophilus, many years passed before cadaver dissections again became common practice (Acar et al 2005).

Galen of Pergamon (129–200 A.D.) was a surgeon of gladiators and gained ample experience treating trauma lesions. In addition to the experience he obtained from clinical observation of his patients, Galen performed anatomical studies on baboons. His detailed descriptions of the cerebral meninges, the corpus callosum, the ventricular system, the pineal glands, and the pituitary as well as his primitive classification of cranial nerves owe much to his dissection of these primates. Galen's studies acquired a paradigmatic status that was maintained for 15 centuries. Despite his incontestable contributions to medicine, the infallibility attributed to him led to the perpetuation of anatomical errors, which survived until questioned by the work of Andreas Vesalius (Goodrich and Flamm 2011; Gordon 1993).

During the Middle Ages, while Europe was still in the age of darkness, the Arab and Byzantine cultures were great centers of intellectual production that maintained their hegemony from 750 to 1200 A.D. Arab scholars preserved the studies of their Greek and Roman predecessors and transcribed and systematized the work of Hippocrates, Herophilus, Galen, and Paulus Aegineta, adding their own observations to it.

The Persian Avicenna (the Westernized form of Abu Ali al-Husayn ibn Abdallah ibn Sina) has become known as "the second doctor" (the first being Aristotle). He was born in Afshana, near Bukhara, in the tenth century. An intellectual with many interests, Avicenna was a big name in his day, writing many books on many different subjects: medicine, philosophy, mathematics, astronomy, and poetry (Pereira 2010).

The noteworthy anatomical descriptions contained in his work *Al-Qanun fi al-Tibb* come from his predecessors' studies and his own astute clinical observations, since anatomical dissection was prohibited. His treatise was very influential and was used as a reference work in Western medical schools until the seventeenth century. Avicenna penned one of the earliest clinical descriptions of epilepsy. His treatise contains surprisingly up-to-date anatomical descriptions of the vertebrae. It correlates the form and size of each vertebra with its location and function, demonstrating a biomechanical understanding that was very advanced for the period (Naderi et al. 2003).

Abulcasis (the Westernized form of Abu al-Qasim Al-Zahrawi) was born in Al-Zahra, near Córdoba, Spain, during the period in which Arabs dominated the Iberian Peninsula. He left us an impressive thirty volumes of material on a variety of medical subjects. The thirtieth volume talks about surgery and is rich in illustrations of surgical instruments. Abulcasis defended the use of cauterization and designed a trepan that did not sink into the cranium, preserving the dura mater. He described ping-pong fractures in detail and observed that they were common in children. In addition, he advocated surgical treatment to alleviate headaches by surgically excising part of the temporal artery, which would probably have been related to headaches caused by temporal arteritis. Abulcasis also described various methods of immobilization for treating vertebral lesions (Nayef et al. 1986).

European medicine resurged at long last in the tenth century, with the creation of the Salerno school. The work of Constantine, the African (1020–1087), who studied in Baghdad, shows the marked influence of Arab medicine. In addition to translating medical manuscripts from Arabic into Latin, Constantino reincorporated cadaver dissection into the study of medicine, although his dissections were performed on pigs, rather than humans. Roger of Salerno treated epileptics with trepanation. He examined cranial trauma victims with the help of what we now know as the Valsalva maneuver, asking his patients to exhale so he could observe the flow of blood or fluid through the edges of cranial fractures (Goodrich and Flamm 2011). Theodoric Borgognoni of Serbia (1205–1298) instituted hygienic measures in treating wounds, avoiding dead space, removing necrotic tissue, and applying wine compresses. William of Saliceto (1210–1277) exchanged Arab cauterization for incisions with a chilled blade. Lanfranchi of Milan (1250–1306) created the first medical school in Italy (Mastronardi and Ferrante 2009). Guy de Chauliac (1300–1368) emphasized the importance of shaving the patient's head before surgery; he developed a classification system for cranial wounds with seven subdivisions and advocated using egg white as a hemostatic (Liu et al. 2003a).

During the sixteenth and seventeenth centuries, the practice of neurosurgery continued to be guided by the appearance of lesions rather than by individual patient's symptoms. Much attention was paid to the anatomical characteristics of the lesions, but the study of neurophysiology and the correlation between location and function had not yet been developed. Leonardo da Vinci (1452–1519) made formidable anatomical studies of the central nervous system. He meticulously analyzed the ventricular system with the help of a wax mold he created. Ambroise Paré (1510–1590), neurosurgeon to the House of Medici, defended debridement of

infected lesions by opening the dura mater and cleaning out the blood and pus. The work of Giacomo Berengario of Capri (1470–1530) contained detailed illustrations of surgical instruments and defended the gravitational draining of intracranial abscesses (Goodrich and Flamm 2011; Liu et al. 2003a). Andreas Vesalius (1514–1564) published his *De Humani Corporis Fabrica*, a book that corrected historical errors that had persisted since the time of Galen, thanks to the lack of studies involving cadaver dissection up to that point. Despite the masterfulness of its world-famous illustrations, the *Fabrica*'s style was verbose and labored, making it difficult to read. In general, bibliographies did not tend to cite his text. The book's illustrations are what mainly get mentioned (Goodrich 1985). Thomas Willis (1621–1625), together with Richard Lower (1631–1691), described the "circle of Willis" that now bears his name.

The eighteenth century begot the work of Percivall Pott (1714–1788) and his description of bone lesions caused by tuberculosis. Pott defended trepanation of trauma lesions in order to alleviate symptoms and not just as a treatment for the fracture itself. Benjamin Bell (1749–1806) described the symptoms of compressive lesions caused by cranial trauma and advocated dealing with them quickly.

Worthy of note not just to the history of neurosurgery, but also to the history of surgery in general, was the development of techniques of anesthesia and of asepsis and antisepsis. Over the course of history, different narcotic substances have been used to alleviate pain, such as coca and opium. But it was not until the nineteenth century that pain control took a great leap forward with the introduction of nitrous oxide by Horace Wells in 1844. William Morton, a Boston dentist, spread the use of ether as an anesthetic agent after 1846. Despite ether's popularity, it irritated the mucosae, and it had to be administered through an uncomfortable system of tubes. James Simpson, a professor of obstetrics in Edinburgh, began a series of studies with different substances until he discovered the power of chloroform. This anesthetic is administered using a cloth soaked in the drug. John Snow perfected the method, when he developed a device for inhaling the drug and regularized the quantity of chloroform needed according to each patient's physical characteristics (Hollingham 2009).

In 1847, Vienna's General Hospital hired Ignaz Philipp Semmelweis. His shrewd observations about puerperal fever have not received the recognition they deserve. Semmelweis believed that medical students' hands perpetuated the illness because they were contaminated with remnants of cadaver material from necropsies. Rigorous guidance about handwashing with calcium chlorate led to a steep drop in mortality rates in maternity wards. Unfortunately, despite all the evidence in their favor, Semmelweis' methods were discounted. Semmelweis was the victim of extreme animosity from his colleagues, and recognition of his work came only posthumously (Céline 1998; Hollingham 2009).

Joseph Lister had more success than Semmelweis. Lister followed with interest the work of Louis Pasteur, which was published between 1857 and 1860. Pasteur annihilated microorganisms through boiling. Since Lister could not apply this method to his patients, he killed microorganisms by applying carbolic acid. This substance was first used in 1865 in curatives applied on an exposed fracture and then came to be used in pulverizers in operating rooms (Hollingham 2009).

After the discovery of anesthesia and of aseptic and antiseptic techniques, the next step was recognizing the signs of cerebral localization. Paul Broca (1824–1880) correlated the third circumvolution of the left cerebral hemisphere with the neural mechanism underlying speech after performing a necropsy on a patient with aphasia. John Hughlings Jackson (1835–1911) performed electrophysiological studies on patients with focal convulsive crises, paving the way for the anatomical localization of lesions. In the absence of any imaging methods, anatomical localization of lesions based on their clinical presentation represented a huge advance for neurosurgery (Black and Black 2001; Liu et al. 2003a).

The Pioneers

Rickman Godlee (1859–1925) described the first tumor resected with a topographic diagnostic basis. The patient had a headache and focal convulsive crises that started in the left face and progressed to the arm and leg. The patient's doctor, John Hughes Bennett, who was not a surgeon, made the diagnosis and localization of the tumor and asked for Godlee's support. The patient's convulsions improved, but he died some time later because of infectious complications (Hollingham 2009).

Victor Alexander Haden Horsley (1857–1916), one of the more talented neurosurgeons in history, performed the first successful laminectomy for spinal tumor in this same period. Full of scientific spirit, Horsley developed a compound based on beeswax for hemostasis of the cranium, performed the first carotid ligation to control a cerebral aneurysm, developed a technique to section the posterior root of the trigeminal nerve to treat trigeminal neuralgia, and studied the location of cerebral functions using an electrostimulation. Horsley worked with Robert Henry Clarke, and together, they created the first stereotactic arch, which they used to stimulate and ablate cerebellar nuclei in animals (Tan and Black 2002).

William Macewen (1848–1924) was a staunch defender of antisepsis. He operated with success on many patients, attributing the satisfactory results he obtained to the combination of a proper topographic diagnosis and the use of rigorous asepsis. Macewen modified Lister's technique and placed pulverized carbolic acid on the operative wound. In his time, he was the neurosurgeon who achieved the most success in treating tumors (Black and Black 2001; Goodrich and Flamm 2011; Liu et al. 2003a).

Harvey William Cushing (1869–1939) is considered one of the most important neurosurgeons of all time. Educated at Yale and Harvard, and trained under William Halsted at Johns Hopkins, Cushing raised neurosurgery to a new level. His meticulous technique and preoccupation with asepsis yielded the highest success rates yet. Cushing defended partial tumor resection in cases where total removal might leave sequelae. He had a great faculty for diagnosis and for correctly locating tumors. He was the author of the first pathological classification of gliomas. Cushing described the symptoms of the illness that today bears his name, and he associated it with the presence of basophils in the pituitary gland. He performed his first transsphenoidal surgery in 1909, using the technique

described by Schloffer, a pioneer in this mode of access for surgery on the pituitary. Cushing later developed his own technique via a sublabial incision. He created various instruments for transsphenoidal use (Black and Black 2001; Goodrich and Flamm 2011; Liu et al. 2003a; Maroon 2005).

Walter Dandy (1886–1946), one of Cushing's pupils, developed the technique of pneumoencephalography, thereby inaugurating the era of using imaging to locate lesions. Dandy proved that neuromas could be resected completely. His many contributions to the study of ventricles included work on colloid cysts of the third ventricle and the identification of the choroid plexus as a source of fluid production (Goodrich 2011; Liu et al. 2003a).

Antônio Caetano de Egas Moniz (1874–1955), a Portuguese surgeon, created the angiography in 1927. Together with Dandy's pneumoventriculography, arteriography allowed for a preoperative evaluation of cerebral lesions in terms of their size, location, and vascularization. Egas Moniz also became famous for developing the frontal lobotomy to control psychiatric illnesses. In 1949, he was awarded the Nobel Prize for what was then considered an advance in the treatment of mental illnesses. Nevertheless, after his death, there were demonstrations demanding that the prize be withdrawn (Hollingham 2011; Black and Black 2001).

Among the innovations from this period, we cannot forget to mention Max Nitze's creation of the first modern endoscope in 1879. In 1922, Dandy first used the endoscope to excise the choroid plexus in a patient with hydrocephalus, but without success. The endoscope continued to be used in neurosurgery for some decades, but with low success rates because of the precariousness of the early apparatuses and inadequate lighting. The creation of the ventriculoperitoneal shunts in 1952, a rapid and effective technique to control hydrocephalus, led to the abandonment of surgical endoscopes. Soon, the simple nature of ventriculoperitoneal shunts would be put to the test. Hyperdrainage, infection, and obstruction were frequent complications that reignited interest in endoscopic treatment (Li et al. 2005). With the birth of fiber optics and the development of new technologies, the endoscope is now used not only to treat hydrocephalus but also to resect intraventricular lesions, to access the base of the cranium, in vertebral column surgeries, and as an aid in conventional microsurgeries.

Decades passed between the creation of the first surgical magnifying glass by the ophthalmologist Edwin Theodor Saemisch in 1876 and the popularization of the microscope in the 1960s. Use of the surgical microscope started in otorhinolaryngology; it was introduced in neurosurgery in 1957. Theodore Kurze used the microsurgical technique to remove a facial schwannoma in a 5-year-old patient, having been inspired by a demonstration video made by otorhinolaryngologist William House. In subsequent years, accessories were created to facilitate its use and to make it more comfortable, such as hydraulic chairs, armrests, and the "diploscope," which allowed a second surgeon to assist (Uluç et al. 2009). Use of the microscope has yielded a better understanding of cerebral anatomy and has allowed less and less invasive surgical techniques to be developed, making it possible to observe in detail structures that are not visible to the naked eye. As a result, hemostasis can be performed more safely and precisely, as can tumor

dissection, with less damage to adjacent tissues. The era of "minimally invasive surgery" had thus begun. Mahmut Gazi Yaşargil was a major contributor to the veritable revolution that followed. Vascular and subarachnoid corridors could be identified for the first time, allowing access to structures that were previously considered unreachable. Yaşargil was named surgeon of the century in 1999 at the Congress of Neurological Surgeons in Boston (Black and Black 2001; Uluç et al. 2009).

The Present

The surgical microscope led to a better understanding of intraoperative anatomy. Other milestones included the introduction of the Malis bipolar forceps, the ultrasonic aspirator, and, more recently, the ultrasound bone cutter. The rise of computed tomography in the 1970s and of imaging magnetic resonance almost a decade later transformed completely the preoperative study of cerebral lesions. For the first time, their morphology could be visualized, and even their functional characteristics could be assessed. Tomography with the administration of a contrast dye can identify breaks in the blood–brain barrier and indicate tumor vascularization grade. Magnetic resonance images have allowed clinicians to estimate the water content of tumors, stage of malignancy, and blood perfusion. Localization systems guided by images have been perfected with the creation of the first arches for tomography-guided stereotaxy and, later, with neuronavigational devices which, in turn, show the location of a surgical instrument in relation to the operative field without having to place a rigid fixation arch on the patient's cranium (Tatagiba et al. 2012). Intraoperative information is correlated with preoperative data, acquired by computed tomography or imaging magnetic resonance, resulting in an image on the monitor that shows, in real time, the surgical instrument's position in relation to the brain or spinal cord (Tatagiba et al. 2012).

In recent years, neurosurgery has made great leaps forward by taking advantage of new and rapidly developing technologies. Compared to the period before the application of these technologies, operations are now being conducted more safely, with less hospitalization time and less discomfort to patients, and with reduced complication rates.

In addition, the surgical environment has changed to accommodate new equipment, taking full advantage of all available space. These changes have borne the concept of integrated operating rooms, which link imaging equipment, such as high-field intraoperative magnetic resonance, to neuronavigational systems. This technology, now widely available, allows brain tumors to be resected more completely (Liu et al. 2003; Roder et al 2012).

Thanks to the perfection of ergonomics, surgeons operate in greater comfort. And now there is hope that, in the near future, we will be able to replace bulky microscopic surgery with a compact and integrated video system, a concept known as video microsurgery (Ebner et al. 2011). In the field of endoscopy, new guided and multidirectional endoscopic prototypes are under development (Ebner et al. 2012a),

as are techniques for perforation and hemostasis assisted by 2-micron fiber lasers for intracranial endoscopic procedures (Ebner et al. 2012b, c). More and more reliable virtual reality devices permit safer planning of surgeries, making an enormous contribution to the training of medical residents, who can now practice many times on simulators before operating on a patient (Elder et al. 2008).

The Future

Robotic assimilation of the brain-computer interface and of nanotechnologies and advances in neuromodulation and stem cells studies are some of the main thrusts for future development.

In the field of robotics, spinal surgery guided by robots is already well established (Devito et al. 2010). Among the current models, *SpineAssist* merits mention; it is used in operations to implant percutaneous pedicle screws for lumbar fusion, significantly improving performance compared to the conventional freehand technique (Devito et al. 2010). Also worthy of mention is the *da Vinci* robotic-assisted laparoscopic anterior lumbar stand-alone interbody fusion procedure via a transperitoneal approach (Beutler et al. 2013). And still in development are assistants for mastoidectomy, a procedure of great precision that requires considerable physical force (Plinkert et al. 2001; Danilchenko et al. 2011).

The period of the last 15 years has brought very promising developments in honing of the brain-computer interface. Science fiction films raised the idea of using thought to move machines (Leuthardt et al. 2006). Today, there are some experimental studies on nonhuman primates and humans in which electrodes are implanted on the cerebral cortex; these electrodes recognize cortical electrical activity and translate signals into actions, such as moving a robotic arm or a suit (exoskeleton) (Leuthardt et al. 2006; Lebedev et al. 2011). In years to come, neurosurgeons will need to remain attentive to brain-computer interface technologies, as well as to the implementation of implantation surgery (Leuthardt et al. 2006). This new era gives the neurosurgeon the opportunity to be active not only in stemming neural tissue damage but also in rehabilitating and recuperating lost functions (Leuthardt et al. 2006).

In the field of nanotechnology, the development of devices with even smaller dimensions will lead to progressively less invasive procedures. Microelectromechanical and nanoelectromechanical systems can act as sensors to identify specific substances, such as markers for tumors or bacterial infections, with infinite possibilities. Nanoparticles offer a means of administering medicine with precision. We will develop systems of biocompatibility to avoid potential immunological reactions to new prostheses (Elder et al. 2008). Finally, as we elucidate the behavior of stem cells, we will be able to harness the way they mature, proliferate, and differentiate to enhance recovery from ischemic brain lesions and degenerative diseases, among other conditions (Elder et al. 2008).

Thus, the future looks promising. Neurosurgery is a constantly evolving field, and it requires practitioners to continually study and extend themselves to absorb

new technologies. We have reached the present state through the hard work of our predecessors. These pioneers opened up the previously unknown territories of anatomy and physiology and discovered the laws of physics and chemistry, guided by their love of knowledge and their humanity. Never before in the history of neurosurgery has science progressed as fast as it has in our era. We must, nonetheless, give a respectful nod to our predecessors, who paved the way for us.

References

Acar, F., Naderi, S., Guvencer, M., Türe, U., & Arda, M. N. (2005). Herophilus of Chalcedon: A pioneer in neuroscience. *Neurosurgery, 56*(4), 861–866.

Andrushko, V. A., & Verano, J. W. (2008). Prehistoric trepanation in the Cuzco region of Peru: A view into an ancient andean practice. *American Journal of Physical Anthropology, 137*, 4–13.

Beutler, W. J. Peppelman, W. C., & Dimarco, L. A. (2013). The da Vinci robotic surgical assisted anterior lumbar interbody fusion: Technical development and case report. *Spine (Phila Pa 1976), 38*(4):356–63.

Black, P. M., & Black, C. T. (2001). History of neurosurgery for intracranial mass lesions. *Neurosurgery Clinics of North America, 12*(1), 1–9.

Carod-Artal, F. J., & Vázquez-Cabrera, C. B. (2004). Paleopatología neurológica em las culturas precolombinas de lacosta y el altiplano andino (II). Historia de las trepanaciones craneales. *Revista de Neurologia, 38*(9), 886–894.

Céline, L. (1998). *The life and works of Semmelweis*. São Paulo: Companhia das Letras.

Danilchenko, A., Balachandran, R., Toennies, J. L., Baron, S., Munske, B., Fitzpatrick, J. M., Withrow, T. J., Webster, R. J., & Labadie, R. F. (2011). Robotic mastoidectomy. *Otology & Neurotology, 32*(1), 11–16.

Devito, D. P., Kaplan, L., Dietl, R., Pfeiffer, M., Horne, D., Silberstein, B., Hardenbrook, M., Kiriyanthan, G., Barzilay, Y., Bruskin, A., Sackerer, D., Alexandrovsky, V., Stüer, C., Burger, R., Maeurer, J., Donald, G. D., Schoenmayr, R., Friedlander, A., Knoller, N., Schmieder, K., Pechlivanis, I., Kim, I. S., Meyer, B., & Shoham, M. (2010). Clinical acceptance and accuracy assessment of spinal implants guided with SpineAssist surgical robot: retrospective study. *Spine (Phila Pa 1976), 35*(24), 2109–2115.

Ebner, F. H., Marquardt, J. S., Hirt, B., Tatagiba, M., & Duffner, F. (2011). Optical requirements on magnification systems for intracranial video microsurgery. *Microsurgery, 31*(7), 559–563.

Ebner, F. H., Nagel, C., Tatagiba, M., & Schuhmann, M. U. (2012a). Efficacy and versatility of the 2-micron continuous wave laser in neuroendoscopic procedures. *Acta Neurochirurgica (Wien), 113*, 143–147.

Ebner, F. H., Hirt, B., Marquardt, J. S., Herlan, S., Tatagiba, M., & Schuhmann, M. U. (2012b). Actual state of EndActive ventricular endoscopy. *Childs Nervous System, 28*(1), 87–91.

Ebner, F. H., Marquardt, J. S., Hirt, B., Honegger, J., Herlan, S., Tatagiba, M., & Schuhmann, M. U. (2012c). Developments in neuroendoscopy: trial of a miniature rigid endoscope with a multidirectional steerable tip camera in the anatomical lab. *Neurosurgical Review, 35*(1), 45–50.

Elder, J. B., Hoh, J. D., Oh, B. C., Heller, C. A., Liu, C. Y., & Apuzzo, M. L. J. (2008). The future of cerebral surgery: A kaleidoscope of opportunities. *Neurosurgery, 62*(3), 1555–1582.

Feldman, R. P., & Goodrich, J. T. (1999). The Edwin Smith Papyrus. *Child's Nervous System, 15*, 281–284.

Goodrich, J. T. (1985). Sixteenth-century renaissance art and anatomy. *Medical Heritage, 1*(4), 280–288.

Goodrich, J. T., & Flamm, E. S. (2011). Historical overview of neurosurgery. In H. R. Winn (Ed.), *Youmans neurological surgery* (6th ed., pp. 3–37). Philadelphia: Elsevier.

Gordon, R. (1993). *The alarming history of medicine*. New York: Martin's Press.
Hollingham, R. (2009). *Blood and guts: A history of surgery*. New York: Thomas Dunne Books.
Lebedev, M. A., Tate, A. J., Hanson, T. L., Li, Z., O'Doherty, J. E., Winans, J. A., Ifft, P. J., Zhuang, K. Z., Fitzsimmons, N. A., Schwarz, D. A., Fuller, A. M., An, J. H., & Nicolelis, M. A. (2011). Future developments in brain-machine interface research. *Clinics, 66*(1), 25–32.
Leuthardt, E. C., Schalk, G., Moran, D., & Ojemann, J. G. (2006). The emerging world of motor neuroprosthetics: A neurosurgical perspective. *Neurosurgery, 56*(1), 1–14.
Li, K. W., Nelson, C., Suk, I., & Jallo, G. I. (2005). Neuroendoscopy: Past, present and future. *Neurosurgical Focus, 19*(6), 1–5.
Liu, C. Y., Spicer, M., & Apuzzo, M. L. J. (2003a). The genesis of neurosurgery and the evolution of the neurosurgical operative environment: Part I – Prehistory to 2003. *Neurosurgery, 52*(1), 3–18.
Liu, C. Y., Spicer, M., & Apuzzo, M. L. J. (2003b). The genesis of neurosurgery and the evolution of the neurosurgical operative environment: Part II – Concepts for future development, 2003 and beyond. *Neurosurgery, 52*(1), 20–33.
Marino, R., & Gonzalles-Portillo, M. (2000). Preconquest Peruvian neurosurgeons: A study of Inca and pre-Columbian trephination and the art of medicine in ancient Peru. *Neurosurgery, 7*, 940–950.
Maroon, J. C. (2005). Skull base surgery: Past, present and future trends. *Neurosurgical Focus, 19*(1), 1–4.
Mastronardi, L., & Ferrante, L. (2009). Neurosurgery in Italy: The past, the present, the future. *Neurosurgical Review, 32*, 381–386.
Naderi, S., Acar, F., Mertol, T., & Arda, M. N. (2003). Functional anatomy of the spine by Avicenna in his eleventh century treatise Al-Qanun Fi Al-Tibb (The Canons of Medicine). *Neurosurgery, 52*(6), 1449–1453.
Nayef, R. F., Al-Rodhan, M. B. B. S., & Fox, J. L. (1986). Al-Zahrawi and Arabian Neurosurgery, 936–1013 AD. *Surgical Neurology, 26*, 92–95.
Panourias, I. G., Skiadas, P. K., Sakas, D. E., & Marketos, S. G. (2005). Hippocrates: A pioneer in the treatment of head injuries. *Neurosurgery, 57*(1), 181–189.
Pereira, R. H. S. (2010). *Avicena: a viagem da alma: uma leitura gnóstico-hermética de Havy ibn Yaqzãn* (2nd ed.). São Paulo: Perspectiva.
Plinkert, P. K., Plinkert, B., Hiller, A., & Stallkamp, J. (2001). Applications for a robot in the lateral skull base. Evaluation of robot-assisted mastoidectomy in an anatomic specimen. *HNO, 49*(7), 514–522. german.
Roder, C., Bender, B., Ritz, R., Honegger, J., Feigl, G., Naegele, T., Tatagiba, M. S., Ernemann, U., & Bisdas, S. (2012, in press). Intraoperative visualization of residual tumor: The role of perfusion-weighted imaging in a high-field intraoperative MR scanner. *Neurosurgery*.
Tan, T., & Black, P. M. (2002). Sir Victor Horsley (1857–1916): Pioneer of neurological surgery. *Neurosurgery, 50*(3), 607–611.
Tatagiba, M., Acioly, M. A., Carvalho, C. H., & Gharabaghi, A. (2012). Advances in neurosurgery. In V. E. Antônio (Ed.), *Neuroscience: Dialogues and intersections* (Chap. 20, 1st ed.). Rio de Janeiro: Editora Rúbio [authors' translation] (in Portuguese).
Uluç, K., Kujoth, G. C., & Baskaya, M. K. (2009). Operating microscopes: Past, present and future. *Neurosurgical Focus, 27*(3), 1–8.

Awake Craniotomies: Burden or Benefit for the Patient?

61

G. C. Feigl, R. Luerding, and M. Milian

Contents

Abstract

An awake craniotomy with intraoperative neuropsychological monitoring and brain mapping is still the gold standard for resections of tumors in or near eloquent areas of the brain. Since it has been shown by several authors that an awake craniotomy is the only way to reliably preserve speech function during tumor resections, the benefits for patients undergoing such a procedure are obvious. However, an awake craniotomy represents an exceptionally stressful situation for a patient, which could possibly lead, similar to an awareness experience during surgery, to long-term psychological sequelae. Therefore the question has to be

G.C. Feigl (✉) • M. Milian
Department of Neurosurgery, University Hospital Tübingen, Tübingen, Germany
e-mail: guenther.feigl@web.de

R. Luerding
Department of Neurology, University of Regensburg Medical Center, Regensburg, Germany

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_63,

raised if it is justified and ethical to submit patients already under a high amount of stress due to the uncertain future and with existential fears to even more strain by operating them awake. However, when balancing the burdens of an awake craniotomy for a patient with the benefits of such a procedure, it becomes obvious that an awake craniotomy and here especially the awake–awake (= continuous awake craniotomy - CAC) method offer the best thinkable balance between an optimal tumor resection and the best possible preservation of cognitive functions. Considering that persistent deficits of essential language and motor function have a much worse effect on a patient's quality of life than rare but treatable posttraumatic stress disorder symptoms, it seems unethical not to offer an awake craniotomy to patients when the location of the tumor justifies such an operation.

Introduction

The first and most important step in any treatment of primary malignant brain tumors is a gross total resection (GTR) (Filippini et al. 2008; Gorlia et al. 2008; Mineo et al. 2007). This however is not always possible because primary brain tumors in most cases infiltrate the surrounding brain tissue. In 1928 Walter Dandy went to an extreme and used hemispherectomies in an attempt to cure patients with malignant primary brain tumors. Yet despite the massive neurological deficits caused by the surgical procedure, these patients suffered recurrences in the opposite hemisphere of the brain. At that time survival of the patient at any prices was the main focus. Even though this seems unethical with today's understanding of neuroethics, the intention of neurosurgeons was the same then as it is today which is to offer the best possible treatment to the patient and save his/her life. The basic idea of Dany to try to remove as much tumor as possible in order to increase the survival of patients with primary brain tumors was correct and has been confirmed in several studies that show that a GTR plays a central part in overall survival of these patients (Berger et al. 1989; Buckner 2003; Stummer et al. 2008; Tait et al. 2007). Nowadays, however, the focus is wider of course when it comes to choosing the best treatment for a patient and also takes the quality of life and preservation of function of a patient into consideration. Nevertheless, the diagnosis of a brain tumor is still a very distressful and threatening situation for a patient. In cases where the tumor is located adjacent to or in an eloquent brain area (i.e., language or motor cortex), a precise intraoperative localization of functional areas is absolutely essential in order to avoid causing deficits during microsurgical tumor removal in eloquent areas of the brain. Even though there are well-known anatomical landmarks to localize functional areas on the cortical surface, they are no longer applicable if functional areas and cerebral tracts are displaced by a space-occupying lesion. Therefore awake craniotomies with intraoperative neuropsychological monitoring and cortical mapping are still the gold standard for tumor resections in or near eloquent cortical areas (Berger et al. 1989; Berger and Ojemann 1992; Duffau 2005; Duffau et al. 2003).

However, an awake craniotomy represents an exceptionally stressful situation for a patient, which could possibly lead, similar to an awareness experience during

surgery, to long-term psychological sequelae (Lennmarken et al. 2002; Leslie and Davidson 2010). However, to this date possible long-term psychological consequences of an awake surgery for neither of the two available protocols (asleep–awake–asleep/awake–awake = continuous awake craniotomy - CAC) are really known. Also, do the benefits for the patients outweigh the burden and stress they are put through when undergoing an awake craniotomy? It has to be considered that postoperative posttraumatic stress disorder (PTSD) symptoms could occur in these neurosurgical patients and certain pretrauma and peritrauma environment factors (e.g., age, gender, pulse rate during mapping, surgery duration) might be associated with psychological symptoms after an awake craniotomy. Furthermore the effect of these symptoms on postsurgical health-related quality of life (HRQoL) has to be taken into consideration when planning an awake surgery for a brain tumor patient. Yet very few studies are available that deal with patients' perceptions of awake craniotomy. Wahab et al. studied the patients' satisfaction with the awake craniotomy procedure and reported that 24 % of the patients experienced some discomfort during surgery, some of which was related to the intraoperative positioning of the patient, and 56 % reported no postoperative pain (Wahab et al. 2011). Whittle et al. reported significant intraoperative pain in 20 % of the patients investigated, and strong anxiety was found in 13 % of the patients (Whittle et al. 2005). In a further study patients were interviewed 1–2 weeks postsurgery (Khu et al. 2010). The results reflected positively on the patients' awake surgery experience, but there were some areas that require improvement, specifically perioperative pain control and postoperative care.

Therefore the question could be raised if it is justified and ethical to submit patients already under a high amount of stress due to the uncertain future and with existential fears to even more strain by operating them awake. For most people it is stressful enough to only think of a dental treatment, but what about an operation on the brain in a conscious state? More so then in a dentist's chair, the patient exposes himself/herself in an awake craniotomy to a situation he/she cannot control which raises fears.

Protocols for Awake Craniotomies

Indications for Awake Craniotomies

The main indications for an awake craniotomy are primary brain tumors in eloquent areas of the brain influencing the speech function. This means any lesion near the two speech centers, Broca and Wernicke and their connecting fiber bundle, the arcuate fascicle, represents an indication for an awake craniotomy. Also patients with Parkinson's disease are generally operated awake so that the correct placement of the electrodes can be tested intraoperatively. The focus of this chapter however is set on awake craniotomies for patients with primary brain tumors. Lesions near the central sulcus no longer represent an indication for an awake craniotomy. The reason is that available electrophysiological monitoring machines allow a very refined localization of the motor area and fiber tracts with cortical and subcortical stimulation on patients even under deep sedation.

The Asleep–Awake–Asleep Protocol

The standard anesthesiological protocol for an awake craniotomy with cortical stimulation and language mapping is the asleep–awake–asleep method (Huncke et al. 1998; Mack et al. 2004; Sarang and Dinsmore 2003; Schulz et al. 2006) named after the three phases the patient is in during the procedure. The patient is "asleep" during the craniotomy, and then sedation is interrupted so the patient is "awake" for cortical stimulation and language testing. After testing is completed, sedation is continued and the patient is again "asleep" for the closure of the skull and skin. Even though this protocol seems to be optimal for patients undergoing awake craniotomies since the patient is left asleep during opening and closure of the skull and is awoken only for language testing, it carries potential risks with respiratory complications and hemodynamic dysregulation being most commonly reported (Danks et al. 2000; Huncke et al. 1998; Keifer et al. 2005; Mack et al. 2004; Manninen et al. 2006; Sarang and Dinsmore 2003; See et al. 2007; Sinha et al. 2007; Skucas and Artru 2006). Furthermore, using the asleep–awake–asleep protocol, patients are always more or less hangover from sedation, leaving them bradyphrenic during neuropsychological testing. This potentially falsifies functional test results creating uncertainty with regard to determining the exact borders of functional areas. Since testing patients intraoperatively is the central part of these procedures, any factors creating an uncertainty during testing should be avoided.

The Awake–Awake Protocol (Continuous Awake Craniotomy/CAC)

In an effort to eliminate potential risks for the patient during awake craniotomies and in order to improve reliability of intraoperative test results which have a direct influence on the surgical outcome of awake craniotomies, the awake–awake protocol has been developed. This protocol is called awake-awake / CAC because using this protocol the patients are not deeply sedated during the procedure and whenever possible no sedatives are used at all. Patients undergoing this procedure receive an intravenous line as well as arterial cannulation (radial artery) for continuous arterial blood pressure monitoring and blood gas analyses. Noninvasive monitoring of vital functions included electrocardiography, a blood pressure cuff, pulsoximetry, and a nasal oxygen mask. Respiratory rate and end-tidal carbon dioxide were measured using a nasal prongs port with capnometry (CO_2 port). Due to the long duration of these procedures, patients receive also a urinary catheter. No airway instrumentation is required but is generally kept within reach of the anesthesiologist in case that a non-self-limiting seizure causing a drop in oxygen saturation (<90 %) occurs.

At least 20 min before the Mayfield head clamp was applied using 0.75 % ropivacaine supplemented with Adrenalin (diluted 1/200,000), scalp nerve blocks (SNB) are placed through a 23 gauge needle (Nguyen et al. 2001a). In order to reduce the psychological stress for the patient in this situation, it is important that a member of the surgical team, either the neuropsychologist or the anesthesiologist, keeps close contact to the patient. This includes not only standing near the patient

but holding his/her hand and talking to him/her and thereby reducing his/her fear because his/her focus is no longer on the surrounding but on the person he/she is talking to. This can be described as a close and individual intraoperative patient guidance and pacing which has been shown to increase the comfort and significantly reduce the stress of patients in such a situation.

Brain Mapping and Neuropsychological Evaluation

Independent of the protocol used, neuropsychological testing and extensive brain mapping with direct cortical stimulation are important to locate functional areas shown on the fMRI scans and confirm their location before the tumor resection is started. Since gliomas generally show an invasive pattern of growths, it is essential to detect functional borders and fiber tracts without compromising the radicality of the resection by stopping the resection too soon. If the tumor is located in or near a functional speech area, neuropsychological testing in combination with subcortical stimulation during the resection helps to detect functional borders and fiber tracts before permanent neurological deficits are caused. Brain mapping, however, can trigger a generalized seizure which is one of the distressful situations for the patient in such a situation. However commonly these types of seizures can be controlled by irrigation of the brain with cold ringer solution (Nguyen et al. 2001a, b; Sartorius and Berger 1998).

For the neuropsychological evaluation during brain mapping and surgery, the patients usually are asked to perform picture naming to identify language areas which are known to be inhibited by stimulation.

Differences in the Anesthesiological Protocols

Management of the airways is a critical part of the standard asleep-awake-asleep protocol with the danger of aspiration due to nausea and vomiting in the phase when the patient wakes up during the surgery for language testing. Various techniques of the airway management are described in literature ranging from leaving the patient breathing spontaneously (Keifer et al. 2005) and securing the airway transnasally with a Magill tube (Schulz et al. 2006) or with a laryngeal mask (Sarang and Dinsmore 2003). In some cases patients are even ventilated using a laryngeal mask (Sarang and Dinsmore 2003) or intubated for the asleep part of the surgery (Huncke et al. 1998). In some centers skull nerve blocks are used for regional pain management during opening of the skin and the craniotomy (Schulz et al. 2006; Sinha et al. 2007). All these manipulations on the patient who is in a "half-awake" state increase his/her feeling of discomfort and helplessness and therefore create stress. A review of recent literature (Table 61.1) reveals that the asleep–awake–asleep protocol is most commonly used during awake craniotomies. These methods have their advantages but also limitations and, more importantly, potential risks are usually not discussed in detail since many publications focus primarily on functional

Table 61.1 A summary of studies on awake craniotomies and reported rates of complication

Author	Number of treated patients	Anesthesiological protocol used	Problems during the wake-up phase or during neuropsychological testing	Extend of tumor resection	Intraoperative seizures	Respiratory complications	Hemodynamic dysregulation	Brain swelling "tight brain"	Other intraoperative complications	New neurological deficits after surgery
Mack et al. (2004)	10	5 asleep–awake–asleep, 5 moderate to conscious sedation	1 (10 %) uncooperative	n. d.	n. d.	2	1 hypotension	1	?	1 accentuation of language difficulties, 2 transient speech problems, 2 transient motor weakness
See et al. (2007)	17	Local anesthesia with moderate to deep sedation	n. d.	n. d.	None	(24 %) Respiratory complications and hypertension				
Low et al. (2007)	20	Local anesthesia with conscious sedation	n. d.	(58 %) Greater 90 % (21 %) greater 80 %	n. d.	?	?	?	6 (30 %) minor anesthetic complications	6 (30 %)
Picht et al. (2006)	25	Asleep–awake–asleep	4 (20 %) restlessness	25 (100 %) GTR	8 (32 %) focal, 2 (8 %) generalized	n. d.	n. d.	1 (4 %)	n.d.	n. d.
Sinha et al. (2007)	42	Local anesthesia with moderate to deep sedation	n. d.	n. d.	4 (9.5 %)	3 (7.1 %) respiratory depression, 2 (4.8 %) desaturation	8 (19 %) hypertension 3 (7.1 %) tachycardia 2 (4.8 %) bradycardia	6 (14.2 %)	7 (16.6 %)	(23.8 %) Transient deficits
Manninen et al. (2006)	50	Asleep–awake–asleep	6 (12 %) uncooperative/ restless	n. d.	4 (8 %)	9 (18 %)	2 hypotension 1 hypertension	2	?	n. d.

Pinsker et al. (2007)	52	25 asleep–awake–asleep, 27 general anesthesia	n. d.	40 (72 %) GTR15 (28 %) STR	None	n. d.	n. d.	n. d.	Temporary neurological deficits 3 speech 2 paresis	14 (26.9 %)
Gupta et al. (2007)	53	27 asleep–awake–asleep 26 general anesthesia	n. d.	(57 %) Greater 90 % (awake group) (73.7 %) greater 90 % (general anesthesia group)	n. d.				?	(23 %) Awake group, (14.8 %) general anesthesia group
Bello et al. (2007)	64	Asleep–awake–asleep	n. d.	(77 %) GTR	(15.6 %)	n. d.	n.d.	n.d.	(6.6 %) Fatigue	32 (50 %)
Keifer et al. (2005)	98	Asleep–awake–asleep	5 (5.1 %) disorientated	n. d.	3 (3 %)	(3 %) Apnea, 2 reintroduction of general anesthesia	None	None	8 nausea 16 headaches	?
Sarang and Dinsmore (2003)	99	Asleep–awake–asleep	2 (2 %) uncooperative	n. d.	5 (5 %)	37	6 hypotension, 4 hypertension	None	24 pain	5
Duffau et al. (2008)	115	Asleep–awake–asleep	n. d.	37 (32.17 %) GTR 59 (51.3 %) STR 19 (16.52 %) PR	n. d.	n.d.	n.d.	n.d.	n.d.	2 (1.7 %) Language deficits
Boulton and Bernstein (2008)	117 biopsies 145 craniotomies	Local anesthesia with moderate to deep sedation	n. d.	n. d.	4 (2.8 %) craniotomy group	n.d.	n.d.	n.d.	n.d.	(5.1 %) Biopsy group, (5.5 %) craniotomy group

and surgical results (Table 61.1). Many authors who do discuss intraoperative complications during awake craniotomies report respiratory complications in up to 37 %, circulatory dysregulation in up to 22 %, brain edema in up to 14 %, and uncooperative or agitated patients in up to 20 % (Danks et al. 2000; Keifer et al. 2005; Mack et al. 2004; Manninen et al. 2006; Picht et al. 2006; See et al. 2007; Sinha et al. 2007). Severe complications with patients waking up uncoordinated and in some cases even pulling themselves out of the Mayfield head clamp inflicting severe injuries on themselves have also been reported. All of these complications are related to sedative agents used during the "asleep" phase of awake craniotomies. Furthermore, these substances have a direct influence on neuropsychological test results; however, intraoperative testing is the prime reason to perform an awake craniotomy in the first place.

The awake–awake / CAC protocol could be safer for patients than other methods used for awake craniotomies. Since there is no drug-related change in the level of consciousness during the entire procedure, it is not surprising that patients are cardiopulmonary stable and that monitoring is generally uneventful for the anesthesiologist. It could be objected that opioids also have a sedative effect. However, the sedative effect of opioids in dosages required for the awake–awake / CAC protocol (approximately 0.1–0.3 mg/h and over a mean period of only 60 min) is by far weaker and therefore not comparable to the sedative effects of barbiturates or other general anesthetics. Also, it is not widely known that several studies specifically analyzing and evaluating the effects of drugs on human memory (Ghoneim 2004a, b) have shown that general anesthetics do impair short-term memory which is highly relevant during intraoperative language and memory testing, while opioids do not show such an effect (Ghoneim 2004b).

Burdens and Benefits of Awake Craniotomies for the Patient

Definition Posttraumatic Stress Disorder (PTSD) and PTSD Symptoms

Since it has been shown by several authors (Berger et al. 1989; Berger and Ojemann 1992; Duffau 2005; Duffau et al. 2003, 2005) that an awake craniotomy is the only way to reliably preserve speech function during tumor resections in or near the speech centers, the benefits for patients undergoing such a procedure are obvious. However, in order to analyze the possible psychological sequelae and their burden on the patient, possible psychological symptoms have to be defined in order to identify them correctly. Based on this information it is then somewhat easier to decide in each individual case whether or not it is ethical to submit a patient to the stress of an awake craniotomy.

According to the definition of the DSM-IV (American Psychiatric Association 1994), PTSD is the development of three main types of characteristic symptoms after an extreme traumatic stress experience (e.g., violation, kidnapping, or the diagnosis of a life-threatening disease). The person's response to the trauma must

involve intense fear, helplessness, or horror (criterion A). The three clusters of symptoms include (i) persistent reexperiencing of the traumatic event (criterion B), (ii) persistent avoidance of stimuli associated with the trauma and numbing of general responsiveness (criterion C), and (iii) persistent symptoms of increased arousal (criterion D). The symptoms must be present for more than 1 month (criterion E) and must cause sufficient distress that leads to occupational or social impairment (criterion F). In the study of Milian et al. (2013) patients were considered to have PTSD symptoms if they reported both having a traumatic event (criterion A) and having at least one of the symptoms from any of the three core clusters (criteria B, C, D). In a previous study it has been shown that positive symptoms for both criterion A and criterion B correctly identified 97 % of PTSD cases. If only single items of the respective criteria (A, B, C, D, or F) are stated positively, patients were considered to show postoperative psychological symptoms; however, these do not fulfill the criteria for a clinically relevant PTSD or PTSD symptoms.

Balancing Burdens and Benefits of an Awake Craniotomy

The risk of affecting the state of arousal of patients during an awake surgery has not yet been quantified in many studies. Since the anesthetist closely observes and protocols the patients' vital parameters such as heart rate, blood pressure, and breathing frequency during an awake craniotomy, he/she has quite sensitive parameters for estimating a change of arousal connected to the awake surgery setting and therefore the possibility to immediately react to it. Therefore the anesthetist has all diagnostic and treatment tools for detecting early signs of stress symptoms and to treat changes of arousal before the level reaches the level of real stress symptoms.

The risk of postoperative psychological sequelae in the sense of posttraumatic stress disorder (PTSD) and the patients' perceptions of awake surgery have so far only been described in very few studies (Milian et al. 2013; Whittle et al. 2005; Khu et al. 2010). Around 12 % of the patients showed PTSD symptoms postoperatively. In one case the symptoms were chronic; in the other case the symptoms resolved within 3 months. However, it is worth mentioning, that no patient that underwent an awake craniotomy fulfilled the criteria for a clinically relevant PTSD after the awake surgery. Physiological parameters such as pulse and blood pressure during the awake craniotomy or the duration of surgery were no predictors for the development of postsurgical psychological symptoms after the procedure. Although PTSD symptoms are rare, isolated psychological symptoms after awake surgery in form of either reexperiencing the event, avoiding of stimuli associated with the event, or symptoms of increased arousal occur to a considerable degree and have a clear negative impact on quality of life (Milian et al. 2013). However, these effects are treatable with a high rate of full recovery (Peris et al. 2011). The rate of recovery of stress symptoms after a suitable therapy exceeds significantly the

effects of neurocognitive training and physiotherapy after postoperative impairment of language and motor functions.

The preservation of language and motor functions is highly significant for the quality of life of a patient. Currently there is no better method available than the awake craniotomy to evaluate speech function during tumor resection. Especially the awake–awake / CAC method offers the best options to test the patient without the influence of sedatives that might falsify intraoperative test results. Furthermore, this protocol significantly reduces the risk for the patient to suffer sedation-related complications such as respiratory complications and hemodynamic dysregulation. Also, potential risks during the intraoperative wake-up phase are avoided since the patient is awake the entire time and intraoperative neuropsychological test results become more reliable. Since the neurosurgeon has to perform the resection to the limit of representations of essential cognitive functions with a maximum of precision, the control of the resection size only by artificial measures such as the fMRI and DTI coordinates bears the incalculable risk of an unjustified confidence in these artificial measures. In awareness of this risk, the neurosurgeon is in most cases far too conservative in estimating the limits of resection (Duffau et al. 2005). This results as a consequence in a smaller tumor volume reduction in comparison to awake resections and therefore in a significantly shorter survival of the patient.

The Protective Effect of Preparing the Patient for an Awake Surgery

Patient selection is a central part in the preparation of an awake craniotomy. The fact that the patient has to play an active role during the entire procedure, especially when using the awake–awake / CAC protocol, might seem frightening to many patients. Besides personality also the ethnic background potentially plays a role in the willingness of a patient to undergo such a surgery. Therefore the neurosurgeon, the anesthesiologist, and the neuropsychologist have to separately talk to the patient and then compare their impressions of the patient and decide whether or not a patient is suitable to undergo an awake surgery. While it is important to openly discuss with the patient the things that can happen during an awake craniotomy, trying to take the patient's fear, it should never be attempted to convince a patient if he/she is not willing or feels not capable of putting himself thru this stressful situation. Patients who are anxious or have a decreased level of motivation generally will not perform well intraoperatively. Therefore, patients who are scheduled for an awake craniotomy have to undergo a thorough preoperative preparation before surgery and should be informed extensively about possible anxiety, pain, noise, seizures, or the inability to name things or speak during language mapping.

During the operation (from skin incision to suture) besides the anesthesiologist, also a neuropsychologist should be present and continuously communicate with the patient. Before performing the craniotomy, the neuropsychologist or the anesthesiologist should try to distract the patient from the stressful situation by trying to associate the noises and phenomenon experienced during the trepanation to

Fig. 61.1 The anesthesiologist keeps close contact to the patient who is operated awake with the awake–awake protocol, placing his/her hand on the patient's shoulder during the trepanation performed by the neurosurgeon in the background of the picture. The anesthesiologist tries to distract the patient from the noise of the drill by trying to associate the phenomenon experienced during the trepanation to a different and positive situation outside the operating room with positive suggestions like "Imagine your are on vacation and your plane is just landing. . .." To keep eye contact between the anesthesiologist and the neurosurgeon, a clear drape is used

a different and positive situation outside the operating room with suggestions like "Imagine your are on vacation and your plane is just landing" (Fig. 61.1). During the surgical procedure, stress for the patient can easily be reduced by engaging him/her into a conversation which at the same time serves as a monitoring of the patient's language function. While it has been described that patients who experienced an awareness during surgery developed PTSD (Lennmarken et al. 2002; Leslie and Davidson 2010), patients who underwent awake craniotomy did not present with similar late psychological sequelae. Therefore, while an unexpected and unprepared awareness during surgery leads to traumatic effects, an adequate preparation seems to have a positive and protective effect for the patient.

Conclusion

Balancing the burdens of an awake craniotomy for a patient against the benefits of such a procedure, it becomes obvious that an awake craniotomy and here especially the awake–awake / CAC method offer the best thinkable balance between an optimal tumor resection and the best possible preservation of cognitive functions. Considering that persistent deficits of essential language and motor function have a much worse effect on a patient's quality of life than rare but treatable posttraumatic stress disorder symptoms, it seems unethical not to offer an awake craniotomy to patients when the location of the tumor justifies such an operation.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Consciousness and Agency
▶ Detecting Levels of Consciousness
▶ Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine
▶ Ethical Implications of Brain Stimulation
▶ Ethics in Neurosurgery
▶ Ethics of Epilepsy Surgery
▶ Ethics of Functional Neurosurgery
▶ Human Brain Research and Ethics
▶ Impact of Brain Interventions on Personal Identity
▶ Informed Consent and the History of Modern Neurosurgery
▶ Neuroimaging Neuroethics: Introduction
▶ Neurosurgery: Past, Present, and Future

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association. Washington. 1-1-1994.

Bello, L., Gallucci, M., Fava, M., Carrabba, G., Giussani, C., Acerbi, F., Baratta, P., Songa, V., Conte, V., Branca, V., Stocchetti, N., Papagno, C., & Gaini, S. M. (2007). Intraoperative subcortical language tract mapping guides surgical removal of gliomas involving speech areas. *Neurosurgery, 60*, 67–80.

Berger, M. S., & Ojemann, G. A. (1992). Intraoperative brain mapping techniques in neuro-oncology. *Stereotactic and Functional Neurosurgery, 58*, 153–161.

Berger, M. S., Kincaid, J., Ojemann, G. A., & Lettich, E. (1989). Brain mapping techniques to maximize resection, safety, and seizure control in children with brain tumors. *Neurosurgery, 25*, 786–792.

Boulton, M., & Bernstein, M. (2008). Outpatient brain tumor surgery: Innovation in surgical neurooncology. *Journal of Neurosurgery, 108*, 649–654.

Buckner, J. C. (2003). Factors influencing survival in high-grade gliomas. *Seminars in Oncology, 30*, 10–14.

Danks, R. A., Aglio, L. S., Gugino, L. D., & Black, P. M. (2000). Craniotomy under local anesthesia and monitored conscious sedation for the resection of tumors involving eloquent cortex. *Journal of Neuro-Oncology, 49*, 131–139.

Duffau, H. (2005). Intraoperative cortico-subcortical stimulations in surgery of low-grade gliomas. *Expert Review of Neurotherapeutics, 5*, 473–485.

Duffau, H., Capelle, L., Denvil, D., Sichez, N., Gatignol, P., Taillandier, L., Lopes, M., Mitchell, M. C., Roche, S., Muller, J. C., Bitar, A., Sichez, J. P., & Van, E. R. (2003). Usefulness of intraoperative electrical subcortical mapping during surgery for low-grade gliomas located within eloquent brain regions: Functional results in a consecutive series of 103 patients. *Journal of Neurosurgery, 98*, 764–778.

Duffau, H., Lopes, M., Arthuis, F., Bitar, A., Sichez, J. P., Van, E. R., & Capelle, L. (2005). Contribution of intraoperative electrical stimulations in surgery of low grade gliomas: A comparative study between two series without (1985–96) and with (1996–2003) functional mapping in the same institution. *Journal of Neurology, Neurosurgery, and Psychiatry, 76*, 845–851.

Duffau, H., Peggy Gatignol, S. T., Mandonnet, E., Capelle, L., & Taillandier, L. (2008). Intraoperative subcortical stimulation mapping of language pathways in a consecutive series of 115 patients with grade II glioma in the left dominant hemisphere. *Journal of Neurosurgery, 109*, 461–471.

Filippini, G., Falcone, C., Boiardi, A., Broggi, G., Bruzzone, M. G., Caldiroli, D., Farina, R., Farinotti, M., Fariselli, L., Finocchiaro, G., Giombini, S., Pollo, B., Savoiardo, M., Solero, C. L., & Valsecchi, M. G. (2008). Prognostic factors for survival in 676 consecutive patients with newly diagnosed primary glioblastoma. *Neuro-Oncology, 10*, 79–87.

Ghoneim, M. M. (2004a). Drugs and human memory (part 1): Clinical, theoretical, and methodologic issues. *Anesthesiology, 100*, 987–1002.

Ghoneim, M. M. (2004b). Drugs and human memory (part 2). Clinical, theoretical, and methodologic issues. *Anesthesiology, 100*, 1277–1297.

Gorlia, T., van den Bent, M. J., Hegi, M. E., Mirimanoff, R. O., Weller, M., Cairncross, J. G., Eisenhauer, E., Belanger, K., Brandes, A. A., Allgeier, A., Lacombe, D., & Stupp, R. (2008). Nomograms for predicting survival of patients with newly diagnosed glioblastoma: Prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. *The Lancet Oncology, 9*, 29–38.

Gupta, D. K., Chandra, P. S., Ojha, B. K., Sharma, B. S., Mahapatra, A. K., & Mehta, V. S. (2007). Awake craniotomy versus surgery under general anesthesia for resection of intrinsic lesions of eloquent cortex–a prospective randomised study. *Clinical Neurology and Neurosurgery, 109*, 335–343.

Huncke, K., Van de Wiele, B., Fried, I., & Rubinstein, E. H. (1998). The asleep-awake-asleep anesthetic technique for intraoperative language mapping. *Neurosurgery, 42*, 1312–1316.

Keifer, J. C., Dentchev, D., Little, K., Warner, D. S., Friedman, A. H., & Borel, C. O. (2005). A retrospective analysis of a remifentanil/propofol general anesthetic for craniotomy before awake functional brain mapping. *Anesthesia and Analgesia, 101*, 502–508, table.

Khu, K. J., Doglietto, F., Radovanovic, I., Taleb, F., Mendelsohn, D., Zadeh, G., & Bernstein, M. (2010). Patients' perceptions of awake and outpatient craniotomy for brain tumor: a qualitative study. *Journal of Neurosurgery, 112*, 1056–1060.

Lennmarken, C., Bildfors, K., Enlund, G., Samuelsson, P., & Sandin, R. (2002). Victims of awareness. *Acta Anaesthesiologica Scandinavica, 46*, 229–231.

Leslie, K., & Davidson, A. J. (2010). Awareness during anesthesia: A problem without solutions? *Minerva Anestesiologica, 76*, 624–628.

Low, D., Ng, I., & Ng, W. H. (2007). Awake craniotomy under local anaesthesia and monitored conscious sedation for resection of brain tumours in eloquent cortex–outcomes in 20 patients. *Annals of the Academy of Medicine, Singapore, 36*, 326–331.

Mack, P. F., Perrine, K., Kobylarz, E., Schwartz, T. H., & Lien, C. A. (2004). Dexmedetomidine and neurocognitive testing in awake craniotomy. *Journal of Neurosurgical Anesthesiology, 16*, 20–25.

Manninen, P. H., Balki, M., Lukitto, K., & Bernstein, M. (2006). Patient satisfaction with awake craniotomy for tumor surgery: A comparison of remifentanil and fentanyl in conjunction with propofol. *Anesthesia and Analgesia, 102*, 237–242.

Milian, M., Luerding, R., Ploppa, A., Decker, K., Psaras, T., Tatagiba, M., Gharabaghi, A., & Feigl, G. C. (2013). "Imagine your neighbour mows the lawn": Does the experience of an awake surgery cause posttraumatic stress disorder symptoms in patients? *Journal of Neurosurgery, 118*, 1288–1295.

Mineo, J. F., Bordron, A., Baroncini, M., Ramirez, C., Maurage, C. A., Blond, S., & Dam-Hieu, P. (2007). Prognosis factors of survival time in patients with glioblastoma multiforme: A multivariate analysis of 340 patients. *Acta Neurochirurgica (Wien), 149*, 245–252.

Nguyen, A., Girard, F., Boudreault, D., Fugere, F., Ruel, M., Moumdjian, R., Bouthilier, A., Caron, J. L., Bojanowski, M. W., & Girard, D. C. (2001a). Scalp nerve blocks decrease the severity of pain after craniotomy. *Anesthesia and Analgesia, 93*, 1272–1276.

Nguyen, A., Girard, F., Boudreault, D., Fugere, F., Ruel, M., Moumdjian, R., Bouthilier, A., Caron, J. L., Bojanowski, M. W., & Girard, D. C. (2001b). Scalp nerve blocks decrease the severity of pain after craniotomy. *Anesthesia and Analgesia, 93*, 1272–1276.

Peris, A., Bonizzoli, M., Iozzelli, D., Migliaccio, M. L., Zagli, G., Bacchereti, A., Debolini, M., Vannini, E., Solaro, M., Balzi, I., Bendoni, E., Bacchi, I., Trevisan, M., Giovannini, V., & Belloni, L. (2011). Early intra-intensive care unit psychological intervention promotes recovery from post traumatic stress disorders, anxiety and depression symptoms in critically ill patients. *Critical Care, 15*, R41.

Picht, T., Kombos, T., Gramm, H. J., Brock, M., & Suess, O. (2006). Multimodal protocol for awake craniotomy in language cortex tumour surgery. *Acta Neurochirurgica (Wien), 148*, 127–137.

Pinsker, M. O., Nabavi, A., & Mehdorn, H. M. (2007). Neuronavigation and resection of lesions located in eloquent brain areas under local anesthesia and neuropsychological-neurophysiological monitoring. *Minimally Invasive Neurosurgery, 50*, 281–284.

Sarang, A., & Dinsmore, J. (2003). Anaesthesia for awake craniotomy–evolution of a technique that facilitates awake neurological testing. *British Journal of Anaesthesia, 90*, 161–165.

Sartorius, C. J., & Berger, M. S. (1998). Rapid termination of intraoperative stimulation-evoked seizures with application of cold Ringer's lactate to the cortex. Technical note. *Journal of Neurosurgery, 88*, 349–351.

Schulz, U., Keh, D., Fritz, G., Barner, C., Kerner, T., Schneider, G. H., Trottenberg, T., Kupsch, A., & Boemke, W. (2006). Asleep-awake-asleep-anaesthetic technique for awake craniotomy. *Anaesthesist, 55*, 585–598.

See, J. J., Lew, T. W., Kwek, T. K., Chin, K. J., Wong, M. F., Liew, Q. Y., Lim, S. H., Ho, H. S., Chan, Y., Loke, G. P., & Yeo, V. S. (2007). Anaesthetic management of awake craniotomy for tumour resection. *Annals of the Academy of Medicine, Singapore, 36*, 319–325.

Sinha, P. K., Koshy, T., Gayatri, P., Smitha, V., Abraham, M., & Rathod, R. C. (2007). Anesthesia for awake craniotomy: A retrospective study. *Neurology India, 55*, 376–381.

Skucas, A. P., & Artru, A. A. (2006). Anesthetic complications of awake craniotomies for epilepsy surgery. *Anesthesia and Analgesia, 102*, 882–887.

Stummer, W., Reulen, H. J., Meinel, T., Pichlmeier, U., Schumacher, W., Tonn, J. C., Rohde, V., Oppel, F., Turowski, B., Woiciechowsky, C., Franz, K., & Pietsch, T. (2008). Extent of resection and survival in glioblastoma multiforme: Identification of and adjustment for bias. *Neurosurgery, 62*, 564–576.

Tait, M. J., Petrik, V., Loosemore, A., Bell, B. A., & Papadopoulos, M. C. (2007). Survival of patients with glioblastoma multiforme has not improved between 1993 and 2004: Analysis of 625 cases. *British Journal of Neurosurgery, 21*, 496–500.

Wahab, S. S., Grundy, P. L., & Weidmann, C. (2011). Patient experience and satisfaction with awake craniotomy for brain tumours. *British Journal of Neurosurgery, 25*, 606–613.

Whittle, I. R., Midgley, S., Georges, H., Pringle, A. M., & Taylor, R. (2005). Patient perceptions of "awake" brain tumour surgery. *Acta Neurochirurgica, 147*, 275–277.

Ethics of Epilepsy Surgery

62

Sabine Rona

Contents

Abstract

One percent of the global population, children as well as adults, are affected by epilepsy. In about one third of patients, seizures cannot be adequately controlled by antiepileptic medication. A significant proportion of these patients, especially of those with focal epilepsy, may benefit from epilepsy surgery. Given the patients' reduced quality of life and the high social and economical burden of uncontrolled seizures, early consideration of epilepsy surgery is recommended. Curative epilepsy surgery procedures aim at abolishing the seizures by removing or disconnecting the area generating the seizures, without causing an unacceptable functional deficit. Palliative procedures aim at reducing the frequency and/or severity of seizures in patients who cannot be rendered seizure-free. Overall, there is a 40–80 % chance of postoperative seizure freedom, with a low risk of perioperative morbidity if the procedure is performed by an experienced

S. Rona
Department of Neurosurgery, University Hospital, Eberhard Karls University, Tübingen, Germany
e-mail: sabine.rona@med.uni-tuebingen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_65,

neurosurgeon. In adults with temporal lobe epilepsy, two randomized controlled trials have provided evidence of the superiority of surgery over medical treatment alone.

To determine whether a patient is a candidate for epilepsy surgery and which procedure is deemed to produce the most benefit, a comprehensive evaluation needs to be performed, taking into account the type of epilepsy and its expected evolution, the individual aims and the psychosocial situation of the patient, the presence or absence of a resectable structural lesion, and the risks associated with the surgery. This chapter examines the ethical challenges related to epilepsy surgery with respect to patient selection, quality of life, and informed consent.

Introduction

Epilepsy is a chronic neurological disease affecting about 0.5–1 % of the general population, children and adults (Hauser et al. 1993). It is characterized by recurrent, unprovoked seizures caused by an inappropriate paroxysmal electrical discharge in the brain. In the majority of patients, the seizures have a focal or regional onset, but most seizures produce activation of a more or less extensive network of cortical and subcortical brain structures. According to the brain areas involved, seizures may manifest as a disturbance of perception, consciousness, or behavior, the maximum expression being a generalized convulsion. About one third of patients with epilepsy are considered drug resistant, meaning that they continue to have seizures despite appropriate pharmacotherapy, or the medication causes unacceptable side effects. It has been shown that the chance of becoming seizure-free with medical therapy alone declines to less than 20 % when a patient has failed adequate trials of 2–3 standard antiepileptic drugs (Kwan and Brodie 2000).

In addition to the impairment of quality of life and social functioning caused by recurrent, unpredictable seizures, uncontrolled epilepsy is associated with significant comorbidities. Mood disorders such as depression or anxiety are four times more frequent than in the general population (LaFrance et al. 2008), and patients often have cognitive deficits due to the underlying disease, medication side effects, or the effect of the seizures themselves. Children with frequent seizures from an early age are impaired in their development (Cross et al. 2006). As they get older, they face increasing psychosocial problems due to the embarrassment and stigma caused by seizures occurring in public, restricted personal autonomy and vocational choices, and the inability to drive motor vehicles. In addition, there is a significantly higher risk of mortality due to injury and sudden death. A recent epidemiological study (Sillanpää and Shinnar 2010) has identified a threefold increase in mortality compared to the general population over an observational period of 40 years.

A substantial proportion of patients with drug-resistant epilepsy, especially of those with focal-onset seizures, may benefit from surgical treatment. In general, two types of surgeries can be distinguished: curative and palliative procedures. Curative procedures aim at achieving seizure freedom by resection or disconnection of the area capable of generating seizures, the so-called epileptogenic zone, without

causing an unacceptable functional deficit. Procedures can be standardized or customized and range from circumscribed cortical excisions, to partial or total resection of an entire lobe of the brain (e.g., anterior temporal lobectomy), to multilobar resections or even hemispherectomy, i.e., removal or functional disconnection of one hemisphere, in patients with severe unihemispheric epilepsy. The amount of brain to be resected or disconnected depends on the nature and localization of the underlying pathology and the presumed extent of the epileptogenic zone. The most frequent pathologies causing drug-resistant seizures amenable to resective epilepsy surgery are low-grade tumors, mesial temporal sclerosis, and malformations of cortical development. In adults, up to 75 % of resections for drug-resistant epilepsy involve the antero-mesial temporal lobe. In children, extra-temporal and multilobar resections are more frequent. Overall, in appropriately selected patients, there is a 40–80 % chance of postoperative seizure freedom, varying according to localization and pathology (Tellez-Zenteno et al. 2005; Spencer and Huh 2008), with a low risk of perioperative morbidity if the surgery is performed by an experienced neurosurgeon. In adults with temporal lobe epilepsy, two randomized controlled trials have provided evidence of the superiority of surgery over medical treatment alone in terms of seizure freedom and quality of life (Wiebe et al. 2001; Engel et al. 2012).

Palliative procedures aim at reducing the frequency and/or severity of seizures in patients who are not considered candidates for curative surgery. This refers to patients with multifocal or bi-hemispheric epilepsies, as well as patients in whom the seizure-onset zone involves eloquent areas. Procedures include the partial resection of the epileptogenic zone in order to spare eloquent cortex, disconnection of cortical areas or white matter tracts to prevent seizure spread such as corpus callosotomy or multiple subpial transections, and electrical stimulation of different targets in the central or peripheral nervous system through an implanted device, with the intent of reducing the excitability of networks thought to play a role in the generation and propagation of seizures.

What distinguishes epilepsy surgery from other neurosurgical interventions to warrant specific ethical considerations? First, it is an elective procedure, and it is (in most cases) irreversible. Therefore, risks and benefits have to be weighed much more carefully than in the case of an emergency intervention undertaken to avert immediate harm. Second, it is a surgical intervention performed with the intent of curing or at least ameliorating a *functional* disorder of the brain. Even though it often entails resection of structural abnormalities, seizures are an electrical phenomenon and the "epileptogenic zone" is an elusive concept. In addition, seizures are generated in potentially functional cortex, meaning that removal of the epileptogenic zone may involve trade-offs in terms of neurological function. Third, many patients who may benefit from epilepsy surgery are legally underage and/or cognitively impaired, which creates additional challenges when obtaining informed consent.

While an in-depth discussion of different ethical theories is beyond the scope of this chapter, the general ethical principles guiding medical professional conduct are considered to be respect for (patient) autonomy, non-maleficence, beneficence, and

justice, with the ultimate aim of achieving the highest good for the patient (Beauchamp and Childress 2009). In the following paragraphs, ethical issues arising in the different stages of the decision-making process will be highlighted.

Candidacy for Epilepsy Surgery

Who is a candidate for epilepsy surgery and how should candidacy be determined? Professional societies concur that every patient with drug-resistant, disabling seizures is a *potential* surgical candidate (Engel et al. 2003; Cross et al. 2006). However, in order to determine whether an individual patient is an *actual* candidate, a comprehensive evaluation needs to be performed, taking into account the type of epilepsy and its expected evolution, the individual aims and the psychosocial situation of the patient, the presence or absence of a resectable structural lesion, and the risks associated with the surgery. This evaluation requires specific expertise, since many epilepsy surgery procedures are not part of standard neurosurgical practice. Given that epilepsy surgery is an elective procedure, timing also becomes an issue. Therefore, while it is the responsibility of the local primary care physicians and specialists (family practitioners, pediatricians, or neurologists) to *initiate* the evaluation for surgical candidacy, the evaluation itself and the final decision whether the patient is likely to benefit from surgical treatment and which type of surgery to recommend should be undertaken in a specialized center with the necessary resources.

Unfortunately, considerable inequity still exists in terms of access to specialized epilepsy centers, with the result that many patients who could benefit from epilepsy surgery are not referred for evaluation. This can often be attributed to informational deficits but in many parts of the world simply to nonexistence or uneven distribution of adequate resources.

The Presurgical Evaluation

The primary purpose of the presurgical evaluation is to identify the so-called "epileptogenic zone," defined as the *minimal* area of cortex that must be resected in order to achieve seizure freedom (Lüders et al. 2006). By definition this includes the seizure-onset zone but also other areas *potentially* able to generate seizures once the current seizure-onset zone has been resected. Therefore, it can be known only *after* the surgery (= when the patient is seizure-free) whether the epileptogenic zone has been inactivated. Since there is no single diagnostic test that allows to determine its extent with certainty, a number of different examinations have to be performed that, taken together, provide an approximation of the epileptogenic zone. In addition to thorough history taking, this includes high-resolution MRI to identify potentially epileptogenic lesions, integrated video-EEG monitoring to delineate the seizure-onset zone, and interictal electrical abnormalities, as well as a neurological examination supplemented by neuropsychological tests and eventually by

functional imaging, to identify functional deficits and indicate eloquent areas that have to be spared during surgery. If the information obtained through noninvasive investigations is not sufficient to make a final decision, or the presumed seizure-onset zone is close to relevant eloquent areas such as language or motor function, invasive video-EEG recordings with implanted subdural or depth electrodes may be necessary to delineate the area to be resected. In this case, as for all invasive procedures, the expected benefit in terms of informational gain has to be weighed carefully against the risk of complications.

Finally, a synthesis of the information obtained has to be created. The more results converge to indicate one circumscribed area, the better the estimate of the epileptogenic zone and the higher the chance of postoperative seizure freedom. If the epileptogenic zone cannot be identified with sufficient certainty or is deemed too extensive or too close to eloquent areas to be removed completely, a cure of the patient's epilepsy becomes unlikely. In this case, a palliative approach may be considered.

The Question of Quality of Life

The goal of epilepsy surgery is to abolish or ameliorate the patients' seizures, but the underlying intent is to improve their quality of life. Beauchamp and Childress (2009) note that it is impossible to determine what will benefit a patient without presupposing some quality-of-life standard and some conception of the life the patient will live after the (surgical) intervention. Therefore, the first question that arises is how to define quality of life and how to measure it. For research purposes, a number of standardized instruments have been developed to evaluate quality of life in general and specific for epilepsy, assessing the patient's subjective perception of the impact of disease and treatment on multiple dimensions of health status, such as physical, psychological, and social functioning, based on the patient's report (Cramer et al. 2002).

Published outcome research has shown that, on average, quality of life as assessed with these tools does indeed improve significantly after epilepsy surgery, in children as well as in adults (Spencer et al. 2007; Seiam et al. 2011; Elliott et al. 2012; Engel et al. 2012; Zupanc et al. 2010). Not surprisingly, the most important *postoperative* determinant of quality of life has been found to be seizure freedom. There is only a limited improvement for patients who are not seizure-free. In this regard it has to be remembered that even for surgical procedures with curative intent, the statistical likelihood of seizure freedom is not 100 % but can be considerably lower according to the accurateness of localization of the epileptogenic zone and the underlying pathology. For palliative procedures, the likelihood of seizure freedom is by definition close to 0 %. In the end, only a minority of patients regret having undergone epilepsy surgery, even if they do not become seizure-free (Macrodimitris et al. 2011). However, while common intuition suggests that any improvement is worth obtaining, and desperate patients as well as healthcare professionals eager to help may be inclined to grasp every straw,

a critical assessment of how much a mere reduction in frequency and/or severity of seizures will actually impact quality of life is paramount when determining the potential benefit for patients with a low chance of seizure freedom.

Outcome research has also shown that seizures, while they tend to dominate patients' preoccupations, are not the only determinant of postoperative quality of life. Not surprisingly, given the high prevalence of neuropsychological impairment and mood disorders in patients with refractory epilepsy, it has been found that the most important *preoperative* determinant is psychological status. This should not lead to the conclusion that patients with psychiatric comorbidities cannot benefit from surgery, and excluding them a priori would go counter the ethical principle of justice, but it has to be kept in mind that these patients require a thorough psychiatric evaluation and that psychological assistance will have to extend into the postoperative period, lest the expected benefit in terms of improvement of quality of life may not manifest.

Last but not least, it should not be forgotten that it can be very difficult to use standardized tools to assess quality of life in patients with reduced mental capabilities who are not able to fill in questionnaires or report on what makes life worth living. For these patients, more "practical" indicators of quality of life, such as injuries through falls or the need to wear a helmet, may need to be applied.

Towards a Patient-Centered Outcome

Meta-analyses of outcome data and guidelines issued by professional associations provide valuable information for the physician seeking to determine the *general* chances and risks of different therapeutic approaches. However, the average patient may be quite different from the individual patient being counseled. While some common determinants exist, quality of life is personal and patients have their own idiosyncratic aims for surgery. The fact that epilepsy surgery is an elective procedure (meaning that the patient voluntarily chooses it with the aim of improving the patient's condition) makes patient satisfaction the most important meter of treatment success. Standardized quality-of-life questionnaires should therefore be supplemented by an in-depth interview with the patient and/or caregivers, in order to really understand the motivation of the patient and his or her social environment (Taylor et al. 2001). The desire to be seizure-free appears quite straightforward, but there are secondary goals such as independence, driving motor vehicles, stopping medications, having more satisfying social contacts, being able to work, finding a partner, etc. Clarifying and discussing individual goals for surgery early in the evaluation process helps to tailor the surgical procedure to the specific needs of the patient and also to uncover unrealistic expectations that will invariably create disappointment after the surgery regardless of its success as defined by "objective" criteria. For example, a patient whose primary motivation for undergoing epilepsy surgery is to drive motor vehicles will likely not be satisfied if he or she has a postoperative visual field defect that prevents driving, even if he or she becomes seizure-free.

Specific Risks of Epilepsy Surgery

In addition to the general surgical risks related to the procedure itself which do not differ from those of other neurosurgical interventions, there are some specific risks of epilepsy surgery that merit consideration, because they have the potential to impact the patients' postoperative quality of life yet are difficult to quantify in terms of either probability or magnitude.

Postoperative Neurological or Neuropsychological Deficits ("Functional Trade-Offs")

As mentioned before, epilepsy surgery often involves resection of functional cortex, which may cause postoperative neurological or neuropsychological deficits such as a worsening of memory in the case of mesial temporal lobe resections or a worsening of motor function in a resection involving primary motor areas. Other neurological or neuropsychological deficits may occur, according to the function represented in the area to be resected. Even in the (hypothetical) case of a certain outcome in terms of seizure freedom, intentionally creating a functional deficit conflicts with the principle of non-maleficence (first, do no harm). It can be ethically justified to override this principle if the expected benefit to the patient outweighs the harm caused by the deficit. There is, however, no general rule how to value the harm caused by an iatrogenic neurological deficit relative to the harm caused by seizures; this is an individual assessment and has to be discussed and agreed upon with the patient. Given the devastating physical and social effects of uncontrolled epilepsy, some patients are willing to sacrifice a considerable amount of function to be relieved of their seizures, but it has to be borne in mind that the seizures are a known entity whereas the impact of the functional deficit may be difficult to anticipate. Outcome research suggests that postoperative neurological deficits only become relevant in terms of quality of life when the patient is not seizure-free (Langfitt et al. 2007). However, given that the chance of postoperative seizure freedom is not 100 %, the risk of becoming a "double loser" has to be considered.

Uncertainties Regarding the Extent of the "Epileptogenic Zone"

The precise extent of the epileptogenic zone is unknown and can only be approximated prior to surgery. This implies a risk of over- or underestimation during the presurgical evaluation process. If the patient is seizure-free after the surgery, the epileptogenic zone is by definition included in the area that has been resected. However, it is possible that a smaller resection may have achieved the same result. This becomes ethically relevant if the resection can cause a neurological deficit. If the extent of the epileptogenic zone is underestimated, the chance of seizure freedom decreases. If it is overestimated, the risk of neurological deficits increases. This dilemma cannot be solved; only its impact can be minimized through the application of professional knowledge. In any case, the potential implications have to be discussed with the patient. Sometimes a two-stage procedure is advocated, undertaking first a smaller resection, with the option of a second, larger, resection if the patient is not seizure-free. This approach minimizes the risk of neurological

deficits but adds the general risks of an additional surgical procedure, as well as the risks due to continuing seizures. Again, this valuation requires a thorough discussion with the patient.

Postoperative Maladaptation ("The Burden of Normality")

It is a declared goal of epilepsy surgery to improve the patient's quality of life by increasing independence and reducing social limitations, but even in the case of successful surgery, the transition to a life without seizures does not always go smoothly. On the one hand, it is not uncommon that patients enter a period of euphoria after surgery which may lead to reckless behavior, creating new health hazards and social conflicts. On the other hand, patients with a long history of epilepsy often have difficulties to discard the "sick" role and adjust to new demands placed upon them (Wilson et al. 2001). The postoperative adaptation process requires time and is prone to disappointment if the "new life" is less satisfactory or takes more time and effort to manifest than has been anticipated. This in turn can influence the perceived success of the procedure.

The Importance of the Multidisciplinary Team

At the end of the evaluation process, an individualized treatment recommendation needs to be formulated. In general terms, potentially curative procedures should take precedence over palliative procedures and the lower the expected benefit in terms of seizure freedom and improvement of quality of life, the lesser the amount of risk that can be tolerated. This may include declining surgery if the perceived benefits do *not* outweigh the risks.

To ensure appropriate decision making, it has to be kept in mind that epilepsy surgery is a multidisciplinary effort. There is an old saying that "to somebody who has a hammer, everything looks like a nail." This also applies to medical professionals in that they have an inherent bias to favor the treatment they are knowledgeable of. In addition, individual physicians or surgeons have personal aims and preferences that may conflict with the patient's best interest. In order to avoid a technical or procedural imperative and to achieve the best possible outcome for the patient, individual expert knowledge has to be balanced by input from other team members with different expertise and points of view. Only when all aspects of the disease and the patient's situation have been considered, a largely unbiased recommendation for the individual patient can be formulated. As a consequence, everybody who has a stake in the presurgical evaluation should also have a word in the decision-making process. In addition to the neurosurgeon(s), this includes neurologists or neuropediatricians, neuroradiologists, neuropsychologists, psychiatrists, as well as social workers in order to facilitate the postoperative transition into a more fulfilling life. Therefore, the final decision whether a patient is a candidate for epilepsy surgery, and which procedure is deemed to produce the most benefit, should be taken by an interdisciplinary patient management committee.

Informed Consent

Once it has been determined that a patient is a candidate for epilepsy surgery and which procedure to recommend, informed consent has to be obtained. The purpose of informed consent is to safeguard patient autonomy in deciding what can be done with their body (or brain, in the case of epilepsy surgery). Autonomy means that an individual is able to act in accordance with a self-chosen plan, free from controlling interference by others as well as from limitations, such as inadequate understanding, that prevent meaningful choice (Beauchamp and Childress 2009). Therefore, the necessary components of legally and ethically valid informed consent are (1) competence, (2) disclosure, (3) understanding, (4) voluntariness, and, finally, (5) consent.

The following refers mainly to patients able to decide for themselves, but its general principles also apply to patients unable to provide legally valid consent and their caregivers or surrogate decision makers. Some specific issues regarding legally incompetent patients, including children, will be discussed at the end of this section.

In order to provide full disclosure, the following elements need to be discussed:

- The patient's epilepsy syndrome and its prognosis
- The recommended surgical procedure – how it is performed, the intended benefits, its likelihood of success, and potential risks
- Possible alternative approaches
- The risks of *not* undertaking the surgery

The physician obtaining consent should have a clear recommendation, be able to elucidate how this recommendation has been reached, and provide competent answers to all possible questions by the patient. This requires first of all that he or she is truly knowledgeable of all the elements stated above. In this regard it may be helpful to have both the neurosurgeon and the neurologist or neuropediatrician participate in the informational exchange.

A structured approach to obtaining patient consent is recommended, but this does not mean the procedure should be overly standardized. Published outcomes provide a useful framework to explain the general chances and risks of the planned procedure, but patients are unique and no two epilepsy surgeries are equal. Interviews with patients have shown clearly that they desire information to be as individualized as possible (Choi et al. 2011). Since outcomes can differ according to experience, the personal experience of the surgeon with this type of surgery should also be included.

When discussing risks of complications with the patient, all generic and specific risks as mentioned previously have to be mentioned. Rare but potentially serious complications should not be neglected. The fact that a serious complication is very rare will become irrelevant for the individual who has to face it. Finally, possible alternative treatments and their chances and risks need to be covered. The risks of *not* undertaking the procedure (i.e., morbidity and mortality due to ongoing seizures) have to be considered as well.

When patients are educated about the implantation of electrodes for diagnostic purposes, it has to be made clear that this does not automatically lead to a resection

and that the final decision can and should be taken only when the results are available. However, essential information about the possible resection should be given beforehand to prevent information overload before the final surgery when decisions may have to be taken under time pressure.

While uncertainties and risks should be clearly stated, being too detailed or presenting too much unstructured information can produce an equation with an unmanageable number of unknowns. Juggling probabilities and weighing uncertainties requires an amount of abstraction not all patients are used to or capable of. Psychological research has shown that human beings are subject to a number of cognitive illusions and biases when elaborating information that can impair informed decision making (Levy 2012). All these biases are accentuated by information overload and time constraints. Therefore, patients tend to extract whatever information they are able (and want) to understand and retain and much depends on how it is presented. In this regard it has to be emphasized that it is the responsibility of the physician obtaining legally binding consent to deliver the pertinent information in a way that the patient is able to understand and to verify what he or she has actually understood; otherwise, there is no informed consent. Ideally, the patient's goals and expectations have been clarified already during the presurgical evaluation, but the preoperative consent procedure is also the last chance to check for unrealistic expectations that can impact patient satisfaction after the surgery. This may include trying to modify the patient's expectations by providing further education and sometimes even imposing unwelcome information if this is deemed to be in the best interest of the patient.

As far as voluntariness is concerned, it has to be kept in mind that patients with drug-resistant epilepsy, like most other patients with chronic illnesses that bring about considerable impairment, are vulnerable to undue influences. In a vulnerable individual, decisional autonomy can be compromised in various ways (Ford 2009). There is a knowledge and status gap as well as a relationship of trust between physicians and patients that invites deferential acceptance of the physicians' recommendations. The lack of alternatives in patients who have exhausted almost all therapeutic options can result in desperation, leading to the acceptance of more risks or lower chances of success than would otherwise be the case. Situational vulnerability may ensue in relation to their social or physical environments. For example, patients may depend on caregivers who have their own interests at stake. Time constraints in decision making, such as when a patient is undergoing invasive monitoring with implanted electrodes which can remain in place only for a limited time, accentuate these vulnerabilities.

While patient vulnerabilities and cognitive biases cannot be neutralized completely, the treating physicians should make every effort to minimize their impact. In addition to awareness of their existence and good communication skills, the most important antidote is *time*. Informed consent is a process that is concluded the day before surgery, but education should begin much earlier. Overall, it is a good idea to address the possibility of surgical therapy already early in the disease, as soon as it becomes clear that drugs do not sufficiently control the seizures. Throughout the presurgical evaluation phase, information from different

sources can then be added on as it emerges, leaving time to digest it, formulate questions, talk to other patients who have undergone the same procedure, obtain a second opinion from experts in the field not directly involved in their care, and eventually adjust expectations.

Patients Who Are Unable to Consent

The level of competence of adult patients, meaning the ability to understand relevant information and make adequate decisions, has to be determined by the treating physician(s). If competence is found to be lacking, a court-appointed surrogate decision maker is needed to provide legally valid consent. For children who have not yet reached the legal age of consent (which may vary according to the country where they live), the decision makers are usually the parents.

Children and adult patients with reduced mental capacities are particularly vulnerable, because their understanding is limited and they are dependent on caregivers. However, the fact that a patient cannot provide legal consent does not mean that he or she is devoid of all decisional capabilities. Mental capacity is a continuum, and the informed consent process should be guided by respect for the patient's individual abilities. Especially children often understand more than adults think and are able to make quite reasonable choices (Alderson 2007). Children as well as incapacitated adults should therefore be involved in the consent process according to their ability, which requires an effort to deliver the relevant information in a way that they can understand. In most cases, at least assent to the procedure can be obtained. Elective surgery should not be performed against the patient's stated will. If a legally incompetent patient refuses to undergo a procedure thought to be in his or her best interest, further education needs to be provided.

A surrogate decision maker is appointed to act in the patient's best interest. Ethical challenges arise whenever it cannot be known what the patient wants, as with small children or adults with very limited mental capabilities. In this case the decision maker has to act in what is *presumed* to be the patient's best interest, which becomes the more difficult the higher the risk or the degree of uncertainty in terms of outcome. Especially parents who have to decide on a surgical procedure involving possible trade-offs in terms of neurological function face a daunting task. On the one hand, early surgery can help to prevent further deterioration through continuing seizures and medication side effects, and the plasticity of the developing brain promises better functional recovery after the surgery. On the other hand, parents are understandably reluctant to take responsibility for inflicting a possibly irreversible neurological deficit on their child, and some would like to postpone this decision to a later date when the child can decide for him- or herself. Another issue that may arise in the case of severely incapacitated patients is whether and to which extent it is legitimate to consider the caregiver's interests. While the patient's quality of life should not be confused with the value of the patient's life for others, it may be argued that alleviating caregiver burden will also benefit the patient.

Whenever the physician obtaining consent has the impression that surrogate decision makers are overwhelmed by their task or conflicts of interest ensue, an ethics consultation should be obtained to provide guidance. Sometimes the court will have to decide.

Cross-References

- ▶ Awake Craniotomies: Burden or Benefit for the Patient?
- ▶ Ethical Implications of Brain Stimulation
- ▶ Ethics in Neurosurgery
- ▶ Ethics of Functional Neurosurgery
- ▶ Free Will, Agency, and the Cultural, Reflexive Brain
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Informed Consent and the History of Modern Neurosurgery
- ▶ Neurosurgery: Past, Present, and Future

References

Alderson, P. (2007). Competent children? Minors' consent to health care treatment and research. *Social Science & Medicine, 65*, 2272–2283.

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics* (6th ed.). New York, Oxford: Oxford University Press.

Choi, H., Pargeon, K., Bausell, R., Wong, J. B., Mendiratta, A., & Bakken, S. (2011). Temporal lobe epilepsy surgery: What do patients want to know? *Epilepsy & Behavior, 22*, 479–482.

Cramer, J. A., Camfield, C., Carpay, H., Helmstaedter, C., Langfitt, J., Malmgren, K., & Wiebe, S. (2002). Principles of health-related quality of life: Assessment in clinical trials. *Epilepsia, 43*, 1084–1095.

Cross, J. H., Jayakar, P., Nordli, D., Delalande, O., Duchowny, M., Wieser, H. G., et al. (2006). Proposed criteria for referral and evaluation of children for epilepsy surgery: Recommendations of the subcommission for pediatric epilepsy surgery. *Epilepsia, 47*, 952–959.

Elliott, I., Kadis, D. S., Lach, L., Olds, J., McCleary, L., Whiting, S., et al. (2012). Quality of life in young adults who underwent resective surgery for epilepsy in childhood. *Epilepsia, 53*, 1577–1586.

Engel, J., Jr., Wiebe, S., French, J., Sperling, M., Williamson, P., Spencer, D., et al. (2003). Practice parameter: Temporal lobe and localized neocortical resections for epilepsy. *Neurology, 60*, 538–547.

Engel, J., Jr., McDermott, M. P., Wiebe, S., Langfitt, J. T., Stern, J. M., Dewar, S., et al. (2012). Early surgical therapy for drug-resistant temporal lobe epilepsy. A randomized trial. *JAMA, 307*, 922–930.

Ford, P. J. (2009). Vulnerable brains: Research ethics and neurosurgical patients. *The Journal of Law, Medicine & Ethics, 37*, 73–82.

Hauser, W. A., Annegers, J. F., & Kurland, L. T. (1993). Incidence of epilepsy and unprovoked seizures in Rochester, Minnesota: 1935–1984. *Epilepsia, 34*, 453–468.

Kwan, P., & Brodie, M. J. (2000). Early identification of refractory epilepsy. *The New England Journal of Medicine, 342*, 314–319.

LaFrance, W. C., Jr., Kanner, A. M., & Hermann, B. (2008). Psychiatric comorbidities in epilepsy. *International Review of Neurobiology, 83*, 347–383.

Langfitt, J. T., Westerveld, M., Hamberger, M. J., Walczak, T. S., Cicchetti, D. V., Berg, A. T., et al. (2007). Worsening of quality of life after epilepsy surgery: Effect of seizures and memory decline. *Neurology, 68*, 1988–1994.
Levy, N. (2012). Forced to be free? Increasing patient autonomy by constraining it. *Journal of Medical Ethics*, BMJ Open Access, published online February 8, 2012.
Lüders, H. O., Najm, I., Nair, D., Widdess-Walsh, P., & Bingaman, W. (2006). The epileptogenic zone: general principles. *Epileptic Disorders, 8*(Suppl. 2), S1–S9.
Macrodimitris, S., Sherman, E. M. S., Williams, T. S., Bigras, S., & Wiebe, S. (2011). Measuring patient satisfaction following epilepsy surgery. *Epilepsia, 52*, 1409–1417.
Seiam, A.-H. R., Dhaliwal, H., & Wiebe, S. (2011). Determinants of quality of life after epilepsy surgery: systematic review and evidence summary. *Epilepsy & Behavior, 21*, 441–445.
Sillanpää, M., & Shinnar, S. (2010). Long-term mortality in childhood-onset epilepsy. *The New England Journal of Medicine, 363*, 2522–2529.
Spencer, S., & Huh, L. (2008). Outcomes of epilepsy surgery in adults and children. Review. *Lancet Neurology, 7*, 525–537.
Spencer, S. S., Berg, A. T., Vickrey, B. G., Sperling, M. R., Bazil, C. W., Haut, S., et al. (2007). Health-related quality of life over time since resective epilepsy surgery. *Annals of Neurology, 62*, 327–334.
Taylor, D. C., McMackin, D., Staunton, H., Delanty, N., & Phillips, J. (2001). Patients' aims for epilepsy surgery: Desires beyond seizure freedom. *Epilepsia, 42*, 629–633.
Tellez-Zenteno, J. F., Dhar, R., & Wiebe, S. (2005). Long-term seizure outcomes following epilepsy surgery: A systematic review and meta-analysis. *Brain, 128*, 1188–1198.
Wiebe, S., Blume, W. T., Girvin, J. P., & Eliasziw, M. (2001). A randomized, controlled trial of surgery for temporal-lobe epilepsy. *The New England Journal of Medicine, 345*, 311–318.
Wilson, S., Bladin, P., & Saling, M. (2001). The "burden of normality": Concepts of adjustment after surgery for seizures. *Journal of Neurology, Neurosurgery and Psychiatry, 70*, 649–656.
Zupanc, M. L., dos Santos Rubio, E. J., Werner, R. R., Schwabe, M. J., Mueller, W. M., & Lew, S. M. (2010). Epilepsy surgery outcomes: Quality of life and seizure control. *Pediatric Neurology, 42*, 12–20.

Ethics of Functional Neurosurgery

63

Robert Bauer and Alireza Gharabaghi

Contents

Abstract

Functional neurosurgery became one of the most dynamic fields in surgery developing from a subdiscipline to a major driver of innovations and novel therapeutic interventions. Despite an inglorious history rooted in unspecific psychosurgical lesioning techniques, functional neurosurgery has evolved to a

R. Bauer (✉) • A. Gharabaghi
Translational and Functional & Restorative Neurosurgery, Department of Neurosurgery, University Hospital Tübingen, Tübingen, Germany

International Centre for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany
e-mail: robert.bauer@cin.uni-tuebingen.de; Alireza.Gharabaghi@med.uni-tuebingen.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_66,

highly precise intervention based on cutting edge imaging, image guidance, and physiological technology offering the last treatment resort for a growing number of neurological and psychiatric indications. Novel devices including closed-loop systems for neuromodulation and prosthetic control promise further new treatment options in the near future.

These dynamic developments are changing the traditional field of related ethical issues significantly. This necessitates that functional neurosurgeons, patients, and society in general will have to deal with new ethical issues in the areas of neuroenhancement, privacy of brain-related information, and patient autonomy regarding control of implanted devices. These issues add to those already inherent to this discipline, e.g., challenges to neurosurgeons from the perspective of professional ethics in the specific context of brain intervention or adequate patient information about the increasing unpredictability of risks and benefits.

Functional neurosurgery will continue to open new doors of modulating brain function; concurrently arising ethical issues need to be addressed by ethicists and physicians jointly and ahead of time.

Introduction

This article on "ethics of functional neurosurgery" is divided in three chapters: core characteristics of functional neurosurgery, specific aspects of functional neurosurgery with ethical relevance, and discussion on general ethical questions.

The current focus of the ethical discussion about functional neurosurgery in literature is the stereotactic implantation of electrodes for deep brain stimulation (DBS). In the last decade more than 130 publications have been written on this topic, in contrast to only six publications on the ethics of functional neurosurgery (according to the data bank PubMed for the search terms "ethic*" $\wedge$ "deep brain stimulation" versus "ethic*" $\wedge$ "functional neurosurgery"). But functional neurosurgery is more than DBS. Although DBS is the most common intervention in functional neurosurgery, it is only one of many techniques used in this field. Thus, the current perspective might mask fundamental anthropological and ethical issues which need to be addressed.

What Is Functional Neurosurgery?

The field of neurosurgery addresses diseases of the central and peripheral nervous system with a large variety of surgical techniques. In the majority of cases, neurosurgical interventions aim at removing pathologies such as brain tumors, vascular lesions, or degenerative tissue of the spinal cord. These "morphological" approaches to diseases of the nervous system have been extended especially in the last two decades by "functional" approaches. Due to neurotechnological developments and a deeper understanding of the neurophysiology of diseases, many new approaches arose modulating "functionally" the nervous system, e.g., in movement disorders

(Elder et al. 2008; Sachdev and Chen 2009; Missios 2007). Hence, these new approaches have to be addressed from an ethical perspective as well.

The term *functional neurosurgery* was coined by Lériche and Wertheimer around 1950, stressing the importance of exploiting neurophysiological knowledge for interventions aimed at changing the dynamics of a system. According to this concept, the "aim and objective of functional neurosurgery are to treat, correct, or balance the functions of the brain that are altered toward either hyperfunctional or hypofunctional states." (Marino 1979). This can result in "positive" or "negative" symptoms and is closely linked to a functional account of disease as exemplified by Boorse. He defined a disease as a "type of internal state which [...] reduces one or more functional abilities below typical efficiency," (Boorse 1977) relating functional abilities to the concepts of reference class and normality: "The reference class is a natural class of organisms of uniform functional design; specifically, an age group of a sex of a species," while "a normal function of a part or process within members of the reference class is a statistically typical contribution by it to their individual survival and reproduction" (Boorse 1977). This approach to the design of neurosurgical interventions necessitates an explicit view of the central and peripheral nervous system as a *dynamical system*. In the same way as the "origin of disequilibrium may be vascular, tumoral, degenerative, or infectious and it may or may not require specific treatment," (Marino 1979) the actual intervention might be targeting a different function located at a different brain area to reestablish normality. In that regard, functional neurosurgery is not a specific technique but a family of methods to modulate functionality with implants in the central or peripheral nervous system.

For this goal, knowledge about the location of functions of the nervous system and their interaction in a network has to be applied with high sensitivity and specificity for each individual patient. Because of these three aspects (localization, system, and subject-level accuracy), presurgical planning and intrasurgical monitoring are of tremendous importance (Ford and Kubu 2005; Kekhia et al. 2011; Borchers et al. 2012; Martino et al. 2011). Functional neurosurgery is therefore highly interested in any advances made in functional brain mapping and neuroimaging which could guide treatment and result in knowledge about biomarkers of disorders (Martin 2012). The fourth important aspect is the increasing understanding of the method of action of these interventions to the nervous system (Cheney et al. 2012; Kringelbach et al. 2010; Dzirasa and Lisanby 2012). This allows functional neurosurgery to improve its tools based on the expected effects of different interventions (Martin 2012; Min et al. 2012). With new technological tools for manipulating functionality and a scientific understanding of the method of action at hand, modern functional neurosurgery has tremendously evolved since the beginnings when therapeutic lesions were applied (Synofzik and Schlaepfer 2008).

Three Main Characteristics

With this in mind, we argue that modern functional neurosurgery is defined by three key characteristics. The first is the concept of *operative neuromodulation*, which is an "interventional field of medicine that alters neuronal signal transmissions by implanted devices, either electrically or chemically, in order to excite, inhibit or

tune the activities of neurons or neural networks to produce therapeutic effects." (Sakas et al. 2007). [S4] The second is the concept of *closed-loop stimulation* which is often also called smart, intelligent, individualized, or on-demand brain stimulation (Benabid et al. 2011; Modolo et al. 2012). The common strategy behind these terms is to monitor online neurochemical (Shah et al. 2010) or neurophysiological activity (Berenyi et al. 2012) for real-time adaption of brain stimulation parameters. The third is the concept of *neural interfacing and prosthetics*, which includes two aspects: the possibility of real-time translation of signals from the nervous system into information which can be utilized for communication, control, or biofeedback (Daly and Wolpaw 2008) and the ability to replace neurological functions such as speech, motor, sensory, hearing, or vision with prosthetic devices at the cellular or systems level (Wang et al. 2010; Stieglitz 2007).

Examples for *operative neuromodulation* are deep brain stimulation (DBS) and epidural motor cortex stimulation (EMCS), which are used in the treatment of Parkinson's disease (Gutiérrez et al. 2009; Fasano et al. 2012), dystonia (Pagni et al. 2008; Vidailhet et al. 2012), or pain (Stadler et al. 2011). Developments of *closed-loop stimulation* are currently evaluated in Parkinson's disease (Tsang et al. 2012; Lee et al. 2009; Priori et al. 2012), pain (Zuo et al. 2012), and epilepsy (Berenyi et al. 2012). Research on *neural interfacing and prosthetics* is addressing speech, e.g., in patients suffering from amyotrophic lateral sclerosis (Brumberg and Guenther 2010) or motor control in patients with spinal cord injury (Bhadra and Chae 2009).

In this context, irreversible therapeutic lesioning is a last resort in functional neurosurgery and results often from insufficient knowledge, technical and financial limitations, or pressing medical needs (Raoul et al. 2009). In contrast, functional neurosurgery aims to keep the anatomy intact and to achieve therapeutic gains by modulating pathological functions. Most other branches of neurosurgery explicitly change anatomy, and obviously, the removal of pathological brain tissue might result in functional improvement as well. Nonetheless, such interventions would be attributed rather to classical general neurosurgery than to functional neurosurgery.

Specific Aspects and Differences to Other Medical Interventions

Although principles from general medical ethics are applicable to functional neurosurgery, this discipline is in many aspects different from other medical specialties. Special attention should be given to the fact that functional neurosurgery is a surgical intervention, usually in the brain, with an infamous history in the era of irreversible brain lesioning. At the same time, while being a last resort for many disorders, it offers unique opportunities for a deeper understanding of brain functioning and disease.

It Is a Surgical Intervention

Any intervention in functional neurosurgery is essentially a surgical one. Therefore the topic of the invasiveness of the intervention is of importance.

Additionally, because the anatomy of every patient is very individual, there is a high need for correct localization. Presurgical mapping based on functional magnetic resonance imaging or transcranial magnetic stimulation becomes increasingly unreliable after the dura has been opened. "Following opening of the dural flap, surgical manipulation, CSF drainage, edema, and issues related to gravity and positioning cause the brain to shift, causing an anatomic discrepancy between the preoperatively acquired images and the surgical field. This displacement is exacerbated by the progression of the surgical procedure." (Kekhia et al. 2011). From this perspective, stereotaxy is only one of the many options available to the functional neurosurgeon to ensure correct targeting. Usually, the neurosurgeon also relies on intraoperative measurements, e.g., spiking activity, local field potentials, or direct cortical stimulation. Because of the brain shift and the potential unreliability of presurgical maps, intraoperative measurements and events have to be interpreted on the spot and a need for adaptation may arise during surgery, also in the case of unplanned complications. This has an influence on how strong a neurosurgeon can rely on evidence-based knowledge for guidance. Often, a surgeon's professional experience and know-how to employ his individual skill and judgment in unique cases must be balanced against a need to generalize knowledge and provide best care for groups of patients (Ford and Henderson 2004). Essentially, functional neurosurgery is a practice with a large heterogeneity between centers (Abosch et al. 2012), and its safety and efficacy relies heavily on the experience, competence, and decisions of the individual surgeon (Kleiner-Fisman et al. 2006; Kirsch and Bernstein 2012). Additionally, many of the functional measurements during the surgery make it necessary that the subject is awake. This presents several unique challenges. Insensitive comments by the staff will be overheard or the patient might misinterpret professional conversation. It is therefore important how patients perceive the surgical team, but this also puts the staff under scrutiny and increases their mental load. Awake surgery can also mean that consent is withdrawn during the intervention (Kirsch and Bernstein 2012).

It Is a Last Resort

Any surgery carries an inherent risk of adverse events of different severity. Adverse events can be attributed to one of three domains. They are either related to the surgery (e.g., transient confusion, hemorrhage, infection, seizures), to the device (e.g., dysfunction, replacement, infection, or migration), or to the modulation (e.g., cognitive, emotional, behavioral, or motor problems) (Kleiner-Fisman et al. 2006). Naturally, surgical risks increase from transcranial (essential noninvasive) to epidural to subdural to subcortical interventions. Complications during surgery can result in severe extension of duration and put a high toll on the stamina of the surgeon. At the same time, this need for individual decision making and the immediate perception of consequences of surgery mean that issues of adequate training, physical ability, professional responsibility, and coping with feelings of guilt present themselves in a special manner in neurosurgery. The challenge for the

surgeon is to trust in one's abilities while being able to perform a critical self-assessment. "All surgeons reach an age when the technical competence and personal stamina necessary to perform surgical procedures may decline. We are responsible to adjust our activity as appropriate when that time occurs." (Umansky et al. 2011). Yet, risks for adverse events during surgery are not only attributable to the surgeon, but are mediated by several factors. Age, neurovascular disorders, other comorbidities, and the psychosocial state of the patient are essential aspects for any consideration of surgery. Therefore evidence-based inclusion and exclusion criteria should be developed to guide risk-benefit assessments. Additionally, there is an inherent, irreducible risk of mortality and permanent impairments. Functional neurosurgery is therefore by many considered to be a last resort treatment when the patient has been shown to be refractory to other treatments. Yet, in many disorders early intervention is being considered. In Parkinson's disease, there is the possibility of a neuroprotective effect and the prevention of psychosocial problems (Schermer 2011); and for amyotrophic lateral sclerosis, the early implantation of communication prosthetics might be necessary to prevent the extinction of thought in a completely locked-in state (Murguialday et al. 2011). Additionally, the different levels of invasiveness should be weighed against the potential benefit.

It Is an Intervention in the Brain

To common understanding, cognitive, sensory, and motor functions are manifested in the nervous system. Functional neurosurgery builds on the assumption that these properties can be localized and modulated. At the same time, the brain has a unique role for most individuals, and also in our society, and is therefore perceived as fundamental for the understanding of personhood. "Recognizing that the brain has a central importance in the organization of patients as persons makes performing brain surgery perceptively different from other types of surgeries." (Ford and Henderson 2004). In parallel, medicalization in the context of a professional interaction is characterized by a low responsibility of the patient for the onset and solution of a problem, but a high expectation to follow the advice of an expert (Brickman et al. 1982). At the same time, neuroscience has a big impact on our concepts of personhood (Farah and Heberlein 2007) and responsibility (Walter 2001), and the topics of medicalization, localization, reification, and exoneration (Fuchs 2006) are ever present. We believe that in many medical specialties, these topics are of reduced presence. But the strong foundation of functional neurosurgery in neuroscience, its high spatial accuracy, the praxis of surgery, i.e., the process of interaction with immediate consequences, and the strong medical primer of a surgery make it next to impossible to ignore these aspects. Therefore the framing of a treatment as "intervening in the brain" (Merkel 2007) or a disease as a "brain disorder" (Leshner 1997) based on functional neuroimaging (Ford and Kubu 2005) changes the perception of the stakes at hand.

It Is an Opportunity for Research

Additionally, any functional intervention in the brain is an opportunity for research and a better understanding of the human nervous system. This research happens as a fundamental part of functional neurosurgery, e.g., when activity in subcortical areas is measured to improve targeting during DBS implantation (Seifried et al. 2012). It can also be a result of postoperative evaluations and help detecting biomarkers of disorders (Bronte-Stewart et al. 2009) or be used for research of functions not directly affected by the disorder (Shibasaki 2012). Balancing the opportunity for research with the best interest of the patients can be a fine line. Patients might feel compelled to satisfy the demands of the clinical personnel or confuse research with a measurement necessary for medical treatment. At the same time, many interventions proposed as treatment are currently still under research. Patients, but also clinicians, can be desperate after unsatisfactory treatments, and the opportunity to take part in a clinical study or explore a novel treatment might feel very compelling in spite of the unknown dangers. Patient autonomy in this regard is therefore no fail-proof guardian against unnecessary or disproportionate risks. On the other side, many neurological disorders such as degenerative disorders or stroke often result in chronic, functional impairments which cannot be treated sufficiently with classical therapeutic approaches. In addition, accidents may result in amputations and losses of sensory organs. Therefore, anyone can get in need for novel functional rehabilitation approaches. Thus, if functional restoration might only be achieved with a neurosurgical intervention and prosthetic approaches, many people will take that risk. This development will be fueled by technological breakthroughs supported by respective funding agencies (Judy 2012).

It Has an Infamous History

Almost no ethical paper in the literature discusses DBS without mentioning the term psychosurgery. Egas Moniz suggested in 1935 the ablation of the frontal cortex as a treatment option for psychiatric disorders which were otherwise treatment resistant. In 1949 he won the Nobel Prize in medicine for this approach. Walter Freeman and James Watts simplified the approach to transorbital frontal lobotomy, on which the infamous story of the ice pick through the eye is being based. From 1945 to 1955 tens of thousands of patients were treated by this crude intervention (Tye et al. 2009). Relatively poor hygienic standards, the conduction with insufficient surgical training, lacking follow-up, and severe side effects resulted in increasing criticism.

Modern functional neurosurgery takes a completely different approach. Many authors see the development of stereotactic procedures as a reaction to the crudeness and high morbidity of lobotomy (Krack et al. 2010). Others see the origin of functional neurosurgery in brain stimulation studies, e.g., in electroconvulsive therapy (ECT) introduced by Ugo Cerletti in 1938 for the treatment of severe psychosis (Sironi 2011), and in reports on intracranial stimulation of the median

forebrain bundle in rats (Appleby and Rabins 2009). Hence, “ablative surgery and electrical stimulation developed in parallel, practically since the introduction of human stereotactic surgery” (Hariz et al. 2010). After high-frequency DBS was introduced and shown to mimic the effect of a lesion, the field of functional neurosurgical interventions changed drastically, as it allowed for adaption of stimulation, reversibility, and reduced morbidity (Benabid et al. 2009a). It should therefore be noted that comparing modern functional neurosurgery to early psychosurgery is wrong in many regards. Modern functional neurosurgery is based on informed consent, clearly defined inclusion and exclusion criteria, clinical decision making by an interdisciplinary team, reversibility and precise targeting, and neuroimaging (Synofzik and Schlaepfer 2008). The comparison of functional neurosurgery to ECT, mind control, or psychosurgery only feeds ungrounded fears (Clausen 2011) and is very unfortunate (Synofzik and Schlaepfer 2008). Yet, the other side of the coin is the hyperoptimistic presentation of DBS in the media (Gilbert and Ovadia 2011).

Ethical Questions

General medical ethics is based on the principles of beneficence, non-maleficence, patient autonomy, and justice (Beauchamp and Childress 2009). These principles also apply to functional neurosurgery. Neuroenhancement and privacy are additional issues that have to be addressed with regard to functional neurosurgery. These issues will be discussed in this chapter.

Beneficence, Non-maleficence, and Cost-Benefit Analysis

Beneficence and its antonym maleficence can be integrated into the concept of a cost-benefit analysis. Due to the large number of interventions possible in functional neurosurgery, a detailed analysis for each intervention is beyond the scope of the present article. As already explained, risks can be related to the surgery, the device, or the modulation. For functional neurosurgery, two aspects are specifically important. First, regular follow-ups are necessary to achieve and maintain the gains of the intervention. This should be factored into any analysis, e.g., by assessing social or familiar support or the distance of the patient’s home to the treating institution for regular follow-ups. Second, a surgery involves not only certain costs, but also risks and uncertainties. Even if a certain treatment would show to be effective in every case, there would still be a risk for a negative outcome due to complications related to the surgery. Therefore, these surgeries cannot be expressed uniquely in cost-benefit analyses, but need a risk assessment as well. As these risks may be related to patient-specific criteria, this means that a reliable cost-benefit analysis can only be performed if definite patient inclusion and exclusion criteria have been established. Yet, humans are notoriously bad at judging and perceiving risks (Lloyd et al. 2001), this is true for both patients and

surgeons. Moreover, quantitative statements may confuse rather than inform those not familiar with these statistical information (Schwartz 2011). It is known that surgeons may suffer from overconfidence in their ability to reduce risks (Kissinger 1998). All these aspects have a direct impact on how patients should be informed for their consent to the intervention.

Autonomy

Every medical intervention needs informed consent, which "implies three basic requests: (1) all medically relevant information about diagnosis and prognosis of a patient's disease, the therapy, its potential risks and alternative therapies must be disclosed. (2) The patient should have the mental capacity to understand his or her situation and the presented information. (3) The patient must not be coerced or compelled, but autonomously decides about a treatment on the basis of the information disclosed." (Skuban et al. 2011). Yet, autonomy in the context of functional neurosurgery faces some special considerations. Many functional neurosurgeries are performed with awake patients. That means, that consent could be revoked at any time during the procedure. Although rare, such cases have been documented. Rules need to be put in place to deal with this situation. Patients may be more likely to suspend a surgery once informed of this option, and surgeons should consider discussing how requests for discontinuation will be handled. Guidelines regarding advanced directives, appropriate intraoperative measures of persuasion, and the possibility of returning to surgery at a later date need to be developed. If the patient would benefit greatly from an intervention, a higher level of convincing might be mandated. Additionally, functional neurosurgery often relies on implanted devices that are in need of ongoing maintenance and professional supervision. Withdrawal of consent by the patient or reduced compliance can result in medical problems and could happen at any time. This withdrawal poses an ethical challenge. In functional neurosurgery, ending treatment will be realized by turning off the device. But in some cases, explantation of the device might be considered – resulting in additional risks. On the other side most of the implantable devices allow for fine-tuning. How much freedom should patients have in changing the parameters of these devices knowing that too much freedom may result in abuse, e.g., self-stimulatory behavior in DBS (Morgan et al. 2006)? At the same time, informed patients' control over the device may reduce side effects of the stimulation and save battery life.

Privacy

With many implanted devices, neuronal activity is constantly monitored during regular activity of the patients. This is commonly the case with electrodes implanted for epileptic diagnostics, but storage is also necessary for many other neural interfaces. Additionally, one might consider other types of data stored in implanted devices, e.g., kinematic data. Already, commercial use of

brain-computer interfaces spawn discussions about side-channel attacks (Martinovic et al. 2012). Data security against malicious intrusion might therefore be topic in the development of safe readouts from implantable devices. There are also non-malicious dangers to privacy. Stored data could be used as evidence in a law trial, and the device could be exploited as a lie detector (of uncertain accuracy). If any information stored in the device is being "disclosed without proper consent, such information could lead to unanticipated insurance, employment, or legal problems for the individual" (Wolpe et al. 2010).

Neuroenhancement

Neuroenhancement has been defined as any "technical intervention aimed at improving some physical or psychological aspect of an individual, but which cannot be categorized as treatment" (Merkel 2007). Currently, to our best knowledge, no functional neurosurgery has been performed worldwide on healthy subjects. It is therefore unknown whether a specific treatment improving functionality in a pathological state might also show efficacy in healthy subjects. That makes it difficult to assess whether neuroenhancement is even feasible. Currently, in the field of *neural prosthetics*, an artificial enhancement is unlikely as any sensory or motor replacement today cannot compete with the original, healthy organ. Yet, additional senses or brain outputs could be an option for enhancement. Brain-computer interfaces used, e.g., for controlling a cursor on the computer screen, already work in healthy subjects with noninvasive recordings (Wolpaw 2007; Silvoni et al. 2011). Invasive recordings would improve the accuracy, speed, and flexibility of the device. Also, noninvasive neuromodulation has been shown to improve a whole range of cognitive and motor functions (Hamilton et al. 2011), from working memory (Polanía et al. 2012; Zaehle et al. 2011) or sensory discrimination (Ragert et al. 2008) to motor performance (Joundi et al. 2012). *Operative neuromodulation*, maybe in combination with *closed-loop stimulation* adapted to the individuals' anatomy and cerebral networks, might further exploit these effects. In the course of a regular treatment, it could happen that a further (or the same) functionality may be improved above the average, an effect coined incidental enhancement by Peter Kramer (Kramer 2009). In depression treatment, mood could be improved above average; in treatment of spinal cord injury, external gadgets could become controlled directly, and in treatment of blindness, additional spectra could be included.

We therefore have no doubt that neuroenhancement with functional neurosurgery is theoretically possible (albeit risky). Therefore, this option might generate a demand. Applications in healthy subjects might be funded by the military. Or wealthy individuals with a transhumanist agenda could start recruiting their own neurosurgical staff. Even when stipulating that "enhancements which cannot at least count as prevention of disease/disability [. . .] should not be included in the sphere of proper medicine as a social system," (Merkel 2007) dealing with privately paid functional neurosurgery may become an issue in the future. Just as plastic surgeons "shifted from reconstructive to cosmetic procedures," (Hamilton et al. 2011)

a similar fate might occur in neurosurgery. Sliding down a slippery slope (McNamee and Edwards 2006) or thanks to a "diagnostic bracket creep," (Kramer 2009) they could become "future providers of neuromodulation technology." (Mendelsohn et al. 2010). From an ethical perspective, that means that it will be important to consider how to deal adequately with such possible developments in advance.

Distinguishing the enhancement of cognitive and motor skills from the modulation of emotion and motivation might be of ethical relevance. Many ascribe the latter functions as more relevant role for the concepts of personal identity (Jotterand and Giordano 2011) or authenticity (Leefman et al. 2011). At the same time, these functions are located more in deeper brain structures (Cardinal et al. 2002); hence, no consistent influence of noninvasive neuromodulation has been shown so far (Hamilton et al. 2011). These targets might be achieved by functional neurosurgery. This means that the issues of personality changes, personhood, authenticity, and personal identity will gain more importance in the assessment of neuroenhancement in this context.

Justice

A main argument against research on functional neurosurgery is that it is too costly and drains funds from established treatments. This concern has, e.g., been expressed against the research of DBS for the treatment of addiction: "The addition of an expensive neurosurgical treatment that costs of the order of US$50 000 (with maintenance costs of approximately US$10 000 every few years) will worsen this situation by utilizing scarce health resources to treat a very small number of patients with the income to pay for it" (Carter and Hall 2011). For many functional neurosurgical interventions, this argument does not hold. DBS, for example, is cost-efficient for addiction (Stephen et al. 2012) and for Parkinson's disease (Valldeoriola et al. 2013). Yet, most DBS surgeries are performed in large teaching hospitals in metropolitan areas (Lad et al. 2010), possibly indicating differences in regional availability of this treatment option. The relative high costs of stimulation devices make it certainly difficult to propagate DBS as a replacement for therapeutic lesions in developing countries (Benabid et al. 2009b), especially when there is already a lack of receiving basic care in these areas (Rosenfeld et al. 2008). At the same time, vulnerable patients from wealthy countries might decide traveling to countries with a lower standard of ethical or scientific scrutiny paying for treatments without any proven benefit (Rosenfeld et al. 2008).

Conclusion

What has already been achieved with functional neurosurgery was a science fiction only a decade ago. Robotic limbs are controlled by thoughts, implanted devices change mood and influence motor skills, and sensors constantly record brain activity and modulate it. We believe that the topics of privacy, autonomy, enhancement, and

individual benefit are already in the focus of functional neurosurgeons. On the other side, while there is awareness that global justice and a fair distribution of health resources are important, these topics receive relatively little attention. Probably this will be one of the issues that will be tackled in future together in the context of studies on the socioeconomic impact of operative neuromodulation, closed-loop stimulation, interfaces, and prosthetics. With the words of William Gibson, "the future is already here – it's just not very evenly distributed."

Cross-References

▶ Awake Craniotomies: Burden or Benefit for the Patient?
▶ Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethical Implications of Brain Stimulation
▶ Ethical Implications of Sensory Prostheses
▶ Ethics in Neurosurgery
▶ Ethics of Epilepsy Surgery
▶ Impact of Brain Interventions on Personal Identity
▶ Mind Reading, Lie Detection, and Privacy
▶ Neuroenhancement
▶ Neuroethics and Identity
▶ Neuroimaging Neuroethics: Introduction
▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives
▶ Research in Neuroenhancement
▶ Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia
▶ Sensory Enhancement

References

Abosch, A., Timmermann, L., Bartley, S., et al. (2012). An international survey of deep brain stimulation procedural steps. *Stereotactic and Functional Neurosurgery, 91*(1), 1–11.

Appleby, B., & Rabins, P. (2009). Ethical considerations in psychiatric surgery. In A. Lozano, P. Gildenberg, & R. Tasker (Eds.), *Textbook of stereotactic and functional neurosurgery* (pp. 2853–2866). Berlin/Heidelberg: Springer. Available at: http://dx.doi.org/10.1007/978-3-540-69960-6_170.

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics*. New York: Oxford University Press.

Benabid, A. L., Chabardes, S., Torres, N., et al. (2009a). Functional neurosurgery for movement disorders: A historical perspective. *Progress in Brain Research, 175*, 379–391.

Benabid, A. L., Chabardes, S., Mitrofanis, J., & Pollak, P. (2009b). Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease. *Lancet Neurology, 8*(1), 67–81.

Benabid, A. L., Costecalde, T., Torres, N., et al. (2011). Deep brain stimulation: BCI at large, where are we going to? *Progress in Brain Research, 194*, 71–82.

Berenyi, A., Belluscio, M., Mao, D., & Buzsaki, G. (2012). Closed-loop control of epilepsy by transcranial electrical stimulation. *Science, 337*(6095), 735–737.

Bhadra, N., & Chae, J. (2009). Implantable neuroprosthetic technology. *NeuroRehabilitation, 25*(1), 69–83.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science, 44*(4), 542–573.

Borchers, S., Himmelbach, M., Logothetis, N., & Karnath, H.-O. (2012). Direct electrical stimulation of human cortex – the gold standard for mapping brain functions? *Nature Reviews Neuroscience, 13*(1), 63–70.

Brickman, P., Rabinowitz, V. C., Karuza, J., et al. (1982). Models of helping and coping. *American Psychologist, 37*, 368–384.

Bronte-Stewart, H., Barberini, C., Koop, M. M., et al. (2009). The STN beta-band profile in Parkinson's disease is stationary and shows prolonged attenuation after deep brain stimulation. *Experimental Neurology, 215*(1), 20–28.

Brumberg, J. S., & Guenther, F. H. (2010). Development of speech prostheses: Current status and recent advances. *Expert Review of Medical Devices, 7*(5), 667–679.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews, 26*(3), 321–352.

Carter, A., & Hall, W. (2011). Proposals to trial deep brain stimulation to treat addiction are premature. *Addiction, 106*, 235–237.

Cheney, P. D., Giffin, D. M., & Van Acker, G. M. 3rd. (2012). Neural hijacking: Action of high-frequency electrical stimulation on cortical circuits. *Neuroscientist*, 1–8.

Clausen, J. (2011). Conceptual and ethical issues with brain-hardware interfaces. *Current Opinion in Psychiatry, 24*(6), 495–501.

Daly, J. J., & Wolpaw, J. R. (2008). Brain–computer interfaces in neurological rehabilitation. *Lancet Neurology, 7*(11), 1032–1043.

Dzirasa, K., & Lisanby, S. H. (2012). How does deep brain stimulation work? *Biological Psychiatry, 72*(11), 892–894.

Elder, J. B., Hoh, D. J., Oh, B. C., et al. (2008). The future of cerebral surgery: A kaleidoscope of opportunities. *Neurosurgery, 62*(6 Suppl 3), 1555–1579; discussion 1579–1582.

Farah, M. J., & Heberlein, A. S. (2007). Personhood and neuroscience: Naturalizing or nihilating? *The American Journal of Bioethics, 7*(1), 37–48.

Fasano, A., Daniele, A., & Albanese, A. (2012). Treatment of motor and non-motor features of Parkinson's disease with deep brain stimulation. *Lancet Neurology, 11*(5), 429–442.

Ford, P., & Henderson, J. (2004). Functional neurosurgical intervention: Neuroethics in the operating room. In J. Illes (Ed.), *Neuroethics*. Oxford: Oxford University Press.

Ford, P., & Kubu, C. (2005). Caution in leaping from functional imaging to functional neurosurgery. *The American Journal of Bioethics, 5*(2), 23–25.

Fuchs, T. (2006). Ethical issues in neuroscience. *Current Opinion in Psychiatry, 19*(6), 600–607.

Gilbert, F., & Ovadia, D. (2011). Deep brain stimulation in the media: Over-optimistic portrayals call for a new strategy involving journalists and scientists in ethical debates. *Frontiers in Integrative Neuroscience, 5*(A16), 1–6.

Gutiérrez, J. C., Seijo, F. J., Alvarez Vega, M. A., et al. (2009). Therapeutic extradural cortical stimulation for Parkinson's disease: Report of six cases and review of the literature. *Clinical Neurology and Neurosurgery, 111*(8), 703–707.

Hamilton, R., Messing, S., & Chatterjee, A. (2011). Rethinking the thinking cap: Ethics of neural enhancement using noninvasive brain stimulation. *Neurology, 76*(2), 187–193.

Hariz, M. I., Blomstedt, P., & Zrinzo, L. (2010). Deep brain stimulation between 1947 and 1987: The untold story. *Neurosurgical Focus, 29*(2), E1.

Jotterand, F., & Giordano, J. (2011). Transcranial magnetic stimulation, deep brain stimulation and personal identity: Ethical questions, and neuroethical approaches for medical practice. *International Review of Psychiatry, 23*(5), 476–485.

Joundi, R. A., Jenkinson, N., Brittain, J.-S., Aziz, T. Z., & Brown, P. (2012). Driving oscillatory activity in the human cortex enhances motor performance. *Current Biology, 22*(5), 403–407.

Judy, J. W. (2012). Neural interfaces for upper-limb prosthesis control: Opportunities to improve long-term reliability. *IEEE Pulse, 3*(2), 57–60.

Kekhia, H., Rigolo, L., Norton, I., & Golby, A. J. (2011). Special surgical considerations for functional brain mapping. *Neurosurgery Clinics of North America, 22*(2), 111–132.

Kirsch, B., & Bernstein, M. (2012). Ethical challenges with awake craniotomy for tumor. *Canadian Journal of Neurological Sciences, 39*(1), 78–82.

Kissinger, J. A. (1998). Overconfidence: A concept analysis. *Nursing Forum, 33*(2), 18–26.

Kleiner-Fisman, G., Herzog, J., Fisman, D. N., et al. (2006). Subthalamic nucleus deep brain stimulation: Summary and meta-analysis of outcomes. *Movement Disorders, 21*(Suppl 14), S290–S304.

Krack, P., Hariz, M. I., Baunez, C., Guridi, J., & Obeso, J. A. (2010). Deep brain stimulation: From neurology to psychiatry? *Trends in Neurosciences, 33*(10), 474–484.

Kramer, P. (2009). Incidental enhancement. In S. Schleidgen, M. Jungert, R. Bauer, & V. Sandow (Eds.), *Human nature and self design* (pp. 155–164). Paderborn: Mentis.

Kringelbach, M. L., Green, A. L., Owen, S. L. F., Schweder, P. M., & Aziz, T. Z. (2010). Sing the mind electric – principles of deep brain stimulation. *European Journal of Neuroscience, 32*, 1070–1079.

Lad, S. P., Kalanithi, P. S., Patil, C. G., et al. (2010). Socioeconomic trends in deep brain stimulation (DBS) surgery. *Neuromodulation, 13*(3), 182–186.

Lee, K. H., Blaha, C. D., Garris, P. A., et al. (2009). Evolution of deep brain stimulation: Human electrometer and smart devices supporting the next generation of therapy. *Neuromodulation: Technology at the Neural Interface, 12*, 85–103.

Leefman, J., Krautter, J., Bauer, R., Tatagiba, M., & Gharabaghi, A. (2011). Die Authentizität modulierter Emotionen bei der Tiefen Hirnstimulation. In L. Kovács (Ed.), *Darwin und die Bioethik Eve-Marie Engels zum 60. Geburtstag*. Orig.-Ausg. Freiburg Br. München: Alber.

Leshner, A. I. (1997). Addiction is a brain disease, and it matters. *Science, 278*(5335), 45–47.

Lloyd, A., Hayes, P., Bell, P. R., & Naylor, A. R. (2001). The role of risk and benefit perception in informed consent for surgery. *Medical Decision Making, 21*(2), 141–149.

Marino, R. (1979). Introduction: Functional neurosurgery as a specialty. In T. Rasmussen & R. Marino (Eds.), *Functional neurosurgery* (pp. 1–6). New York: Raven Press.

Martin, J. H. (2012). Systems neurobiology of restorative neurology and future directions for repair of the damaged motor systems. *Clinical Neurology and Neurosurgery, 114*(5), 515–523.

Martino, J., Honma, S. M., Findlay, A. M., et al. (2011). Resting functional connectivity in patients with brain tumors in eloquent areas. *Annals of Neurology, 69*(3), 521–532.

Martinovic, I., Davies, D., & Frank, M. u. a. (2012). On the feasibility of side-channel attacks with brain-computer interfaces. In *21st USENIX Security Symposium*. USENIX Association.

McNamee, M. J., & Edwards, S. D. (2006). Transhumanism, medical technology and slippery slopes. *Journal of Medical Ethics, 32*(9), 513–518.

Mendelsohn, D., Lipsman, N., & Bernstein, M. (2010). Neurosurgeons' perspectives on psychosurgery and neuroenhancement: A qualitative study at one center. *Journal of Neurosurgery, 113*(6), 1212–1218.

Merkel, R. (2007). *Intervening in the brain changing psyche and society*. Berlin/New York: Springer.

Min, H.-K., Hwang, S.-C., Marsh, M. P., et al. (2012). Deep brain stimulation induces BOLD activation in motor and non-motor networks: An fMRI comparison study of STN and EN/GPi DBS in large animals. *NeuroImage, 63*(3), 1408–1420.

Missios, S. (2007). Hippocrates, Galen, and the uses of trepanation in the ancient classical world. *Neurosurgical Focus, 23*(1), 1–9.

Modolo, J., Beuter, A., Thomas, A. W., & Legros, A. (2012). Using "smart stimulators" to treat Parkinson's disease: Re-engineering neurostimulation devices. *Frontiers in Computational Neuroscience, 6*, 69.

Morgan, J. C., diDonato, c J., Iyer, S. S., et al. (2006). Self-stimulatory behavior associated with deep brain stimulation in Parkinson's disease. *Movement Disorders, 21*(2), 283–285.

Murguialday, A. R., Hill, J., Bensch, M., et al. (2011). Transition from the locked in to the completely locked-in state: A physiological analysis. *Clinical Neurophysiology, 122*(5), 925–933.

Pagni, C. A., Albanese, A., Bentivoglio, A., et al. (2008). Results by motor cortex stimulation in treatment of focal dystonia, Parkinson's disease and post-ictal spasticity. The experience of the Italian Study Group of the Italian Neurosurgical Society. *Acta Neurochirurgica Supplement, 101*, 13–21.

Polanía, R., Nitsche, M. A., Korman, C., Batsikadze, G., & Paulus, W. (2012). The importance of timing in segregated theta phase-coupling for cognitive performance. *Current Biology, 22*(14), 1314–1318.

Priori, A., Foffani, G., Rossi, L., & Marceglia, S. (2012). Adaptive deep brain stimulation (aDBS) controlled by local field potential oscillations. *Experimental Neurology, 245*, 77–86.

Ragert, P., Vandermeeren, Y., Camus, M., & Cohen, L. G. (2008). Improvement of spatial tactile acuity by transcranial direct current stimulation. *Clinical Neurophysiology, 119*(4), 805–811.

Raoul, S., Leduc, D., Deligny, C., & Lajat, Y. (2009). Therapeutic lesions through chronically implanted deep brain stimulation electrodes. In A. M. Lozano, P. L. Gildenberg, & R. R. Tasker (Eds.), *Textbook of stereotactic and functional neurosurgery* (pp. 1427–1442). Berlin/Heidelberg: Springer.

Rosenfeld, J. V., Bandopadhayay, P., Goldschlager, T., & Brown, D. J. (2008). The ethics of the treatment of spinal cord injury: Stem cell transplants, motor neuroprosthetics, and social equity. *Topics in Spinal Cord Injury Rehabilitation, 14*(1), 76–88.

Sachdev, P. S., & Chen, X. (2009). Neurosurgical treatment of mood disorders: Traditional psychosurgery and the advent of deep brain stimulation. *Current Opinion in Psychiatry, 22*(1), 25–31. doi:10.1097/YCO.0b013e32831c8475.

Sakas, D. E., Panourias, I. G., Simpson, B. A., & Krames, E. S. (2007). An introduction to operative neuromodulation and functional neuroprosthetics, the new frontiers of clinical neuroscience and biotechnology. *Acta Neurochirurgica Supplement, 97*(Pt 1), 3–10.

Schermer, M. (2011). Ethical issues in deep brain stimulation. *Frontiers in Integrative Neuroscience*, *5*(A17), 1–5.

Schwartz, P. H. (2011). Questioning the quantitative imperative: Decision aids, prevention, and the ethics of disclosure. *The Hastings Center Report, 41*(2), 30–39.

Seifried, C., Weise, L., Hartmann, R., et al. (2012). Intraoperative microelectrode recording for the delineation of subthalamic nucleus topography in Parkinson's disease. *Brain Stimulation, 5*(3), 378–387.

Shah, R. S., Chang, S.-Y., Min, H.-K., et al. (2010). Deep brain stimulation: Technology at the cutting edge. *Journal of Clinical Neurology, 6*(4), 167–182.

Shibasaki, H. (2012). Cortical activities associated with voluntary movements and involuntary movements. *Clinical Neurophysiology, 123*(2), 229–243.

Silvoni, S., Ramos-Murguialday, A., Cavinato, M., et al. (2011). Brain-computer interface in stroke: A review of progress. *Clinical EEG and Neuroscience, 42*(4), 245–252.

Sironi, V. A. (2011). Origin and evolution of deep brain stimulation. *Frontiers in Integrative Neuroscience*, *5*(A42), 1–5.

Skuban, T., Hardenacke, K., Woopen, C., & Kuhn, J. (2011). Informed consent in deep brain stimulation – ethical considerations in a stress field of pride and prejudice. *Frontiers in Integrative Neuroscience, 5*(A7), 1–2.

Stadler, J. A., 3rd, Ellens, D. J., & Rosenow, J. M. (2011). Deep brain stimulation and motor cortical stimulation for neuropathic pain. *Current Pain and Headache Reports, 15*(1), 8–13.

Stephen, J. H., Halpern, C. H., Barrios, C. J., et al. (2012). Deep brain stimulation compared with methadone maintenance for the treatment of heroin dependence: A threshold and cost-effectiveness analysis. *Addiction, 107*(3), 624–634.

Stieglitz, T. (2007). Restoration of neurological functions by neuroprosthetic technologies: Future prospects and trends towards micro-, nano-, and biohybrid systems. *Acta Neurochirurgica Supplement, 97*(Pt 1), 435–442.

Synofzik, M., & Schlaepfer, T. E. (2008). Stimulating personality: Ethical criteria for deep brain stimulation in psychiatric patients and for enhancement purposes. *Biotechnology Journal, 3*(12), 1511–1520.

Tsang, E. W., Hamani, C., Moro, E., et al. (2012). Subthalamic deep brain stimulation at individualized frequencies for Parkinson disease. *Neurology, 78*(24), 1930–1938.

Tye, S. J., Frye, M. A., & Lee, K. H. (2009). Disrupting disordered neurocircuitry: Treating refractory psychiatric illness with neuromodulation. *Mayo Clinic Proceedings, 84*(6), 522–532.

Umansky, F., Black, P. L., DiRocco, C., et al. (2011). Statement of ethics in neurosurgery of the world federation of neurosurgical societies. *World Neurosurgery, 76*(3–4), 239–247.

Valldeoriola, F., Puig-Junoy, J., & Puig-Peiró, R. (2013). Cost analysis of the treatments for patients with advanced Parkinson's disease: SCOPE study. *Journal of Medical Economics, 16*(2), 191–201.

Vidailhet, M., Jutras, M. -F., Grabli, D., & Roze, E. (2012). Deep brain stimulation for dystonia. *Journal of Neurology Neurosurgery Psychiatry*, 1–14.

Walter, H. (2001). *Neurophilosophy of free will: From libertarian illusions to a concept of natural autonomy*. Cambridge, MA: MIT Press.

Wang, W., Collinger, J. L., Perez, M. A., et al. (2010). Neural interface technology for rehabilitation: Exploiting and promoting neuroplasticity. *Physical Medicine and Rehabilitation Clinics of North America, 21*(1), 157–178.

Wolpaw, J. R. (2007). Brain-computer interfaces as new brain output pathways. *Journal of Physiology (London), 579*(Pt 3), 613–619.

Wolpe, P. R., Foster, K. R., & Langleben, D. D. (2010). Emerging neurotechnologies for lie-detection: Promises and perils. *The American Journal of Bioethics, 10*(10), 40–48.

Zaehle, T., Sandmann, P., Thorne, J. D., Jäncke, L., & Herrmann, C. S. (2011). Transcranial direct current stimulation of the prefrontal cortex modulates working memory performance: Combined behavioural and electrophysiological evidence. *BMC Neuroscience, 12*, 2.

Zuo, C., Yang, X., Wang, Y., et al. (2012). A digital wireless system for closed-loop inhibition of nociceptive signals. *Journal of Neural Engineering, 9*(5), 056010.

Section XIV

Addiction and Neuroethics

What Is Addiction Neuroethics?

64

Adrian Carter and Wayne D. Hall

Contents

Abstract

Drug use and addiction are significant problems facing most societies. Neuroscience promises to reduce the incidence and harms of drug use through the development of more effective treatments targeted at changes in the brain produced by chronic drug use and by identifying persons most likely to develop harmful drug use and preventing them from becoming addicted. By locating the source of addictive behavior in the brain, neuroscience may be seen to account for some addicted individuals' criminal behavior and justify the use of coercive interventions to treat their addiction. Neuropharmacological interventions used to treat drug addiction may also be co-opted by healthy individuals to enhance their normal cognition. In this chapter we introduce the reader to *Addiction Neuroethics*, the ethical issues raised by proposed and potential applications of neuroscience research on addiction.

Addiction is a central topic in *Neuroethics*. Drug use and addiction are major problems facing most societies and one of the largest causes of preventable disease

A. Carter (✉)
The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia
e-mail: adrian.carter@uq.edu.au

W.D. Hall
The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia

Queensland Brain Institute, The University of Queensland, St Lucia, QLD, Australia
e-mail: w.hall@uq.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_67,

burden worldwide accounting for over 10 % of the overall burden of disease in Europe and developed countries such as Australia, Canada, and the USA (Begg et al. 2007; ONDCP 2004; Rehm et al. 2005). Neuroscience research has begun to describe in great detail the impact of chronic use of drugs on reward systems in the brain that leads to addiction. Addiction also raises many neuroethical challenges, arguably more so than many other psychiatric or neurological disorders, because of the complex intersection of medical, social, and legal responses to a strongly socially disapproved form of behavior.

Addiction is typically defined as a persistent pattern of drug use in the face of harms caused to the drug user and others in their social environment, such as family, friends, workmates, and neighbors. There has been a long running opposition between moral and medical explanations of this pattern of drug use. A moral model describes addiction as a largely voluntary behavior in which people choose to engage (Heyman 2009), and hence, drug users who commit criminal offences should be prosecuted and imprisoned if found guilty. A medical model of addiction, by contrast, recognizes that while many people can use drugs without losing control over their use, a minority will develop a mental or physical disorder – an addiction – that undermines their ability to control their drug use and requires treatment if the sufferer is to become and remain abstinent (Leshner 1997).

Neuroscience research on addiction is seen as supporting a version of the medical model that explains addictive behavior in terms of changes in brain processes produced by chronic drug use (Koob and Le Moal 2006). If accepted, such causal accounts provide an alternative to the view that the use of addictive drugs is always a matter of individual choice. Proponents of the brain disease model of addiction believe that it will radically benefit social and public health policies towards addicted persons by providing more effective, humane, and ethical treatment of addicted persons (Dackis and O'Brien 2005; McLellan et al. 2000). These benefits will include:

- Less reliance on expensive punitive policies such as imprisonment
- More funding of addiction research and treatment services
- Greater access to treatment for those who are addicted
- Increased coverage of addiction treatment by insurance companies
- Greater investment in addiction treatment, research, and development
- Decreased stigmatization of addicted persons to reduce social isolation and increase treatment seeking
- A more rational approach to drug policy that reflects the impact of specific drugs upon the brain

Social scientists have highlighted potentially less welcome effects of neuroscience models of addiction on social and public health policy (Campbell 2012; Carter and Hall 2012; Midanik 2004). They argue that neuroscience research overemphasizes biomedical approaches to the treatment of addiction at the expense of psychosocial policies that aim not only to treat addiction but also to reduce drug use in society.

The five papers contained in this section of the *handbook* analyze some of the ethical and legal issues raised by neuroscience research on addiction and its current and future applications to treating and preventing addiction and developing social policies towards drug use.

Anne Lingford-Hughes and Liam Nestor provide an overview of neuroscience research on addiction (▶ Chap. 65, "Neuroscience Perspectives on Addiction: Overview"). They outline the types of research evidence that are used to support the claim that addiction can be usefully thought of as a "brain disease." This includes insights derived from influential animal models of human addiction, such as estimating the addictive potential of new psychoactive substances and identifying how these drugs act within the brain to change brain structure and function. They summarize recent neuroimaging studies of addicted persons which suggest that the findings from animal models of addiction apply in addicted humans.

▶ Chap. 66, "Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response" addresses the ethical implications of claims that neuroscience research will soon enable us to identify persons who are at greatest risk of developing an addiction. This research is often promoted as a way of allowing the development of social policies that will more effectively prevent addiction in persons at higher risk and match addicted individuals to the forms of treatment that are most likely to benefit them. Such proposals, should they prove successful, would raise ethical concerns, such as privacy and third party uses of predictive information and the use of preventive interventions without consent. In this chapter, Wayne Hall, Adrian Carter, and Murat Yücel review recent neuroimaging studies that are claimed to have identified markers of brain activity in persons who are most susceptible to developing an addiction. They begin by reviewing genetic research conducted with similar aims to see what might be gleaned from this experience about the likely utility of neuroprediction. They outline the technical and ethical challenges that remain to be addressed before neuroimaging technologies can be used either for preventive purposes or to match addicted persons to the form of treatment that is most likely to assist them to become and remain abstinent.

In ▶ Chap. 67, "Ethical Issues in the Treatment of Addiction," Benjamin Capps and colleagues examine the ethical issues that may arise in the application of knowledge from addiction neurobiology to the treatment of addiction. Their chapter covers ethical issues raised by the treatment of opioid and other forms of drug addiction using substitute medications; the use of drug vaccines and depot forms of antagonists to treat addiction; the use of invasive treatments to directly modify brain function, such as deep brain stimulation and neurosurgery; and the use of varying degrees of coercion, including legal coercion by the courts, to encourage addicted persons to enter addiction treatment.

▶ Chap. 68, "Drug Addiction and Criminal Responsibility" considers the implications of addiction neuroscience for the way in which the criminal law treats addicted offenders. Jeanette Kennett, Nicole Vincent, and Anke Snoek examine the bearing of neuroscience evidence on the assessment of the responsibility of addicted persons for their drug use and crime engaged in to finance it. They then consider the implications of this assessment for the use of imprisonment and legal coercion in the treatment of addicted offenders. They accept that the majority of addicted individuals do not suffer from a brain disease that overwhelms their autonomy but argue nonetheless that some severely addicted individuals do develop a brain disease that undermines their ability to control their drug use and reduces

their responsibility for some criminal acts committed to facilitate their addictive behavior. They provide results from interviews from a group of addicted individuals to support their thesis.

► Chap. 69, "Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues" examines possible regulatory approaches to the use of pharmaceutical drugs that are claimed to enhance the cognitive functioning in normal individuals, that is, persons who do not have cognitive impairment. Jayne Lucke and colleagues analyze current approaches to the regulation of psychoactive drugs that are used for nontherapeutic reasons, such as alcohol and tobacco, pharmaceuticals, and illicit drugs, and the possible implications of neuroscience research on drug addiction for the regulation of putatively neuroenhancing pharmaceuticals.

Cross-References

► Drug Addiction and Criminal Responsibility
► Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response
► Ethical Issues in the Treatment of Addiction
► Neuroscience Perspectives on Addiction: Overview
► Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Begg, S., Vos, T., Barker, B., et al. (2007). *The burden of disease and injury in Australia 2003*. Canberra: Australian Institute of Health and Welfare.

Campbell, N. D. (2012). Medicalization and biomedicalization: Does the diseasing of addiction fit the frame? In J. Netherland (Ed.), *Critical perspectives on addiction* (Advances in medical sociology, Vol. 14, pp. 3–25). Bradford: Emerald Group.

Carter, A., & Hall, W. (2012). *Addiction neuroethics: The promises and perils of neuroscience research on addiction*. London: Cambridge University Press.

Dackis, C., & O'Brien, C. (2005). Neurobiology of addiction: Treatment and public policy ramifications. *Nature Neuroscience, 8*, 1431–1436.

Heyman, G. (2009). *Addiction: A disorder of choice*. Cambridge: Harvard University Press.

Koob, G. F., & Le Moal, M. (2006). *Neurobiology of addiction*. New York: Academic.

Leshner, A. I. (1997). Addiction is a brain disease, and it matters. *Science, 278*, 45–47.

McLellan, A. T., Lewis, D. C., O'Brien, C. P., et al. (2000). Drug dependence, a chronic medical illness: Implications for treatment, insurance, and outcomes evaluation. *Journal of the American Medical Association, 284*, 1689–1695.

Midanik, L. T. (2004). Biomedicalization and alcohol studies: Implications for policy. *Journal of Public Health Policy, 25*, 211–228.

ONDCP. (2004). *The economic costs of drug abuse in the United States*. Washington, DC: Office of National Drug Control Policy.

Rehm, J., Room, R., van den Brink, W., et al. (2005). Problematic drug use and drug use disorders in EU countries and Norway: An overview of the epidemiology. *European Neuropsychopharmacology, 15*, 389–397.

Neuroscience Perspectives on Addiction: Overview

65

Anne Lingford-Hughes and Liam Nestor

Contents

Abstract

Substance addiction can be a chronic relapsing disorder. While different drugs of addiction have different primary molecular targets, it has been demonstrated that many share the common action of being able to increase dopamine within hard-wired reward circuitry. While this effect is widely conceived as a primary factor driving initial drug use, long-term adaptations within this hard-wired neural circuitry underlie the transition from drug use to drug dependence. Significantly, these neuroadaptations are responsible for triggering recurrent drug relapse in people recovering from addiction, even when following periods of long-term abstinence. While there is no animal model of addiction that can fully emulate the human condition, some animal models do permit the investigation of specific elements of drug addiction, particularly those involving the reward system and its role in drug-seeking behavior. Neuroimaging methods now also permit us to test hypotheses of addiction derived from such animal models, allowing the field

A. Lingford-Hughes (✉) • L. Nestor
Centre for Neuropsychopharmacology, Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK
e-mail: anne.lingford-hughes@imperial.ac.uk; anne.lingford-hughes@ic.ac.uk; liam.nestor@imperial.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_68,

of neuroscience to examine neural components of drug abuse and dependence in humans. These neuroimaging procedures permit neuroscientists to test hypotheses in humans at different stages of the addiction cycle, particularly with a view to developing better treatments.

Introduction

Substance addiction is often defined as a chronic, relapsing disorder characterized by (1) compulsion to seek and take the substance, (2) loss of control in limiting substance intake, and (3) the emergence of a negative emotional state (e.g., dysphoria, anxiety, irritability) reflecting a motivational withdrawal syndrome when access to the substance is prevented (defined as Substance Dependence by the 4th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) of the American Psychiatric Association). Individuals meeting such criteria are those commonly described in clinical and neuroscience literature; however, many people do not enter such a chronic relapsing pattern and recover without treatment. Although drugs of addiction (e.g., alcohol, amphetamines, cocaine, nicotine) have varying pharmacological profiles, their ability to activate the mesocorticolimbic system is known to mediate their acute reinforcing effects (Koob and Volkow 2010). With chronic drug use, however, it is proposed that long-term neuroadaptations within this same mesocorticolimbic circuitry underlie the transition from drug use to drug dependence and relapse to drug use during abstinence.

Animal models have been critical in developing theories regarding the evolution of addiction. The positive-reinforcing effects of drugs of abuse (i.e., their ability to induce a conscious feeling of pleasure) have been widely conceived as a primary factor behind continued drug use and eventual drug dependence (Koob and Volkow 2010). Both positive and negative reinforcement theories additionally provide some insight into both the initiation (i.e., pleasure) and maintenance (i.e., withdrawal avoidance) of compulsive drug use (Cami and Farre 2003; Koob and Volkow 2010). Such theories, however, are unable to account for the resumption of drug-seeking and drug-taking behaviors (i.e., relapse) following protracted periods of abstinence. Chronic drug use has been proposed to result in a pathological shift in the hedonic set point of the drug user (Koob and Le Moal 1997). This state of dysregulation within evolutionary hard-wired brain reward systems is proposed to ultimately lead to a loss of control over drug use; that is, drug abuse produces a disequilibrium within brain reward circuitry that cannot be biologically maintained without using the substance of addiction.

An alternative theory is that addiction involves the development of "incentive sensitization" (Berridge 2009; Berridge and Robinson 1998; Berridge et al. 2009; Robinson and Berridge 2000). The chronic use of drugs is theorized to produce alterations in neural systems involved in the motivation and reward for natural appetitive reinforcers (e.g., food, water). Drug abuse is thought to induce a hypersensitive state to the drug and drug-associated stimuli (e.g., people, places,

and objects associated with the substance). This ultimately leads to a shift from drug "liking" to drug "wanting," whereby there is an ensuing compulsion to seek out and use the substance at any cost.

Additional theories view the persistent nature of addictive behaviors as ingrained drug habits in the form of aberrant stimulus response learning (Everitt and Robbins 2005; Volkow et al. 2006a; Wise 2002) and alterations in prefrontal cortical activity that reduce behavioral control and decision-making skills (Bechara 2005; Goldstein and Volkow 2002). Although each of these theories contributes its own unique perspective to the evolution of addiction, it is likely that there is significant overlap in these concepts. Furthermore, in humans, initial drug use probably involves a complex interaction between various components (i.e., biological, sociological, economic, legal, and cultural) that, in some individuals, produce substance addiction.

Pharmacology of Major Drugs of Dependence

Understanding the pharmacology of drugs of abuse is vital to improving the prevention and treatment of addiction (Nutt and Lingford-Hughes 2008; Lingford-Hughes et al. 2010). Drugs of abuse are commonly classified into major categories (Feltenstein and See 2008) that include narcotics (e.g., opiates such as heroin), cannabinoids (e.g., marijuana), depressants (e.g., alcohol), and stimulants (e.g., nicotine, amphetamines, cocaine). While these substances all produce feelings of pleasure and relieve negative emotional states, they also possess highly diverse behavioral effects due to their varied neuropharmacological profiles in the brain. The mesocorticolimbic dopaminergic system is a key target for all substances of abuse. Modulation of the dopaminergic system may occur directly, as in the case of stimulants that block the dopamine transporter (DAT) in the nucleus accumbens (e.g., cocaine) or stimulate dopamine release (e.g., amphetamine). It may also be indirect, by increasing dopaminergic neuronal firing via disinhibition of inhibitory gamma amino butyric acid (GABA) interneurons in the ventral tegmental area (e.g., alcohol, opiates, nicotine) (see Everitt and Robbins 2005).

A key modulator of mesolimbic dopaminergic function is the endogenous endorphin system. It is the major target for the opioid drugs such as morphine and heroin, and it has long been implicated in processes such as interpersonal bonding (e.g., mother-child), love, and reward. Opiates reduce anxiety and induce euphoria and sedation (Heishman et al. 2000; Hill and Zacny 2000) by activating opioid receptors of which there are three subtypes: the mu opioid receptor (mOR), kappa opioid receptor (kOR), and delta opioid receptor (dOR). The activity of opiates at the mOR subtype underlies their abuse potential. The mOR is located in a variety of brain regions, including the cerebral cortex, thalamus, hippocampus, locus coeruleus, ventral tegmental area (VTA), nucleus accumbens (NAcc)/ventral striatum (VS), and the amygdala. Opiates mediate their reinforcing effects directly in the NAcc at the mOR and indirectly through mOR inhibition of GABA function

on dopaminergic cells in the VTA, thereby increasing firing of dopamine VTA projections to the NAcc. Recent findings from human brain imaging studies suggest that addiction is associated with alterations within the endorphin system and that the craving, distress, and dysphoria found in early alcohol and drug abstinence are associated with alterations in mOR (see section "Neuroimaging Studies in Addiction" below).

Emerging research also suggests that the kOR subtype may play a role in addiction, particularly the experience of negative emotional states during drug withdrawal (Bruijnzeel 2009). Stimulation of kOR inhibits dopamine release in the striatum, with chronic administration of drugs of abuse shown to increase the release of dynorphin in this region. This suggests that the chronic abuse of substances (not just opiates) may have an enduring effect at the kOR subtype.

Cannabinoids induce feelings of euphoria, disinhibition, relaxation, and analgesia (Curran et al. 2002). Delta-9-tetrahydrocannabinol (or THC) is the principal psychoactive constituent of cannabis and exerts its central effects via the cannabinoid 1 receptor (CB_1). CB_1 receptors are highly expressed in the cerebral cortex, hippocampus, striatum, amygdala, and cerebellum (Herkenham et al. 1991; Tsou et al. 1998). CB_1 receptors are also located in the VTA and NAcc, where they modulate dopaminergic firing (D'Souza et al. 2008; Huestis et al. 2007; Hunault et al. 2008). The presence of the CB_1 receptor in the hippocampus is believed to underlie the effects of cannabinoids on memory.

Alcohol also induces euphoria, relaxation, and disinhibition, while reducing stress and anxiety (Koob 2004). The reinforcing effects of alcohol likely arise from its interaction with numerous neurotransmitter systems in the brain. Two key systems involved are opiate and GABA (the brain's major inhibitory neurotransmitter). Alcohol increases endogenous endorphin release and modulates $GABA_A$ and $GABA_B$ receptors to increase dopamine levels (Koob 2004; Sullivan et al. 2011; Tang et al. 2003). Stimulation of GABA results in anxiolysis and sedation, which are major drivers for alcohol abuse in humans (Sieghart 2006). Tolerance to alcohol involves adaptations in the GABA system as well as excitatory glutamatergic *N*-methyl-D-aspartate (NMDA) receptors, making the GABA system less responsive and increasing NDMA receptor activity. In the absence of alcohol, these adaptions lead to a hyperexcitable brain, resulting in signs and symptoms of alcohol withdrawal that include tremor, fits, and delirium tremens.

While tobacco use is not associated with significant psychological and social impairment typical of other types of addiction, it is the leading cause of preventable death in developed countries (Benowitz 2008; Mathers and Loncar 2006; Peto et al. 1996). Nicotine is the main addictive component in cigarettes (Benowitz 2009; Gray et al. 1996; Mansvelder and McGehee 2002) and acts on nicotinic acetylcholine receptors (nAChRs) in the brain, including those in the VTA, that modulate dopaminergic cell firing (Grady et al. 2010; Klink et al. 2001; Wooltorton et al. 2003; Drenan et al. 2010; Mameli-Engvall et al. 2006). The high-affinity $\alpha4\beta2$ subunit of the nAChR appears to be crucial to the positive-reinforcing and cognitive-enhancing effects of nicotine (Lippiello et al. 2006; Patterson et al. 2009).

Positron emission tomography (PET) studies in humans have demonstrated that smoking produces dopamine release in the VS (Brody et al. 2004). Furthermore, research in humans, investigating the effects of smoking on α4β2 nAChRs, has shown that smoking a full cigarette results in more than 88 % α4β2 subunit receptor occupancy, an effect which is accompanied by a significant reduction in cigarette craving (Brody et al. 2006). These PET research findings corroborate the value of medications that specifically target the α4β2 subunit of nAChRs (Gonzales et al. 2006; Jorenby et al. 2006) in reducing relapse in smokers attempting to quit.

Psychostimulants, such as cocaine and amphetamines, directly increase the concentration of dopamine in mesocorticolimbic brain regions (Kuhar et al. 1991; Wise 1996). Cocaine is a reuptake inhibitor that binds to the presynaptic dopamine transporter (DAT) (Amara and Kuhar 1993; Woolverton and Johnson 1992) that moves dopamine from the synapse back into presynaptic nerve terminals. By blocking the DAT, cocaine inhibits the reuptake of dopamine, increasing dopaminergic levels and amplifying its reinforcing effects. Amphetamines also block the DAT but directly trigger dopamine release as well (Rudnick and Clark 1993). The increase in dopamine produced by psychostimulants correlates with the resulting "high" that people experience (Volkow et al. 1999). While the pharmacological effects of psychostimulants also increase levels of serotonin and norepinephrine in the brain (Howell and Kimmel 2008; Rudnick and Clark 1993; Sora et al. 2009), it is primarily their effects on dopamine at the NAcc that underlie their abuse potential. Ecstasy or MDMA, another amphetamine-type drug, primarily targets the serotonergic system by blocking serotonin reuptake, producing a different set of euphoric experiences.

Animal Models of Addiction

There are a variety of increasingly sophisticated animal models that have provided invaluable insights into both the neurobiology of addiction and the pharmacological actions of drugs of abuse (Feltenstein and See 2008; Heidbreder 2011; Yahyavi-Firouz-Abadi and See 2009). Models that have particularly served to elucidate important neurobehavioral mechanisms in addiction include intracranial self-stimulation (ICSS), conditioned place preference (CPP), behavioral sensitization, and self-administration paradigms. Furthermore, these models, particularly those involving self-administration, have also proved beneficial in examining the neurobiology and neuropharmacology of drug relapse.

Knowledge about brain regions important for reward originally began with research in rodents by Olds and Milner, who found that rats would expend a great deal of effort to electrically stimulate areas of the brain that form part of a reward circuit (Olds and Milner 1954). Additional evidence suggested that the rewarding effects of ICSS activated a dopamine projection from the VTA to the NAcc, via a pathway known as the medial forebrain bundle (Heimer and Van Hoesen 2006; Wise 2005). Drugs of abuse have been observed to decrease ICSS thresholds; that

is, the reinforcing properties of addictive drugs reduce the amount of brain stimulation required by the animal. Animal research has also revealed that the more addictive a substance is, the greater its ability to reduce the ICSS threshold. This model can be used to evaluate the abuse potential of different drugs. The ICSS model has also served as a unique experimental tool to assess alterations in the basal hedonic state of an animal following chronic drug exposure. In contrast, withdrawal from all major drugs of abuse produces an *increase* in ICSS thresholds; that is, the effects of drug withdrawal on reward circuitry increase the amount of brain stimulation required by the animal to overcome this state.

The CPP model utilizes the classical (Pavlovian) conditioning paradigm in which an animal learns associations between a conditioned stimulus (CS) and unconditioned stimulus (UCS). A CS (e.g., neutral object) is repeatedly paired with an UCS (e.g., drug), until the CS on its own comes to elicit the same response as the UCS. In the CPP model, an animal is exposed to an apparatus consisting of two neutral environments. These environments can differ in terms of a number of stimulus modalities, including color, texture, odor, and lighting (Bardo and Bevins 2000). One environment is paired with drug administration (CS+), while the other is paired with the administration of a control substance, usually saline (CS-). After a number of conditioning sessions, the animal (now in a drug-free state) is permitted free access to the environments of the apparatus, during which their preference for the two environments is measured (e.g., by frequency of entry into and the time spent in the environments). In accordance with the principles of classical conditioning, because the drug condition has reinforcing effects, the animal shows a significant preference for the drug-paired (CS+), over the saline-paired (CS−), environment.

Experimental studies show that various drugs of abuse (e.g., amphetamines, cocaine, heroin, nicotine) typically induce CPP for the drug-paired environment (Pastor et al. 2012; Sticht et al. 2010; Thorn et al. 2012), suggesting a role for classical conditioning in the acquisition of drug use behavior. The CPP model, however, possesses a number of limitations, including the method of drug administration (e.g., experimenter administered) that fails to model human drug use (i.e., self-administration), the potential confound of novelty on the day of testing, difficulties in generating dose–response curves, and the model being limited to use in rodents.

Behavioral sensitization involves a progressive increase in the motor stimulatory effects of a drug with repeated and intermittent administration. The development of behavioral sensitization has been hypothesized to represent the shift from drug "liking" to drug "wanting" that has been hypothesized to underlie compulsive drug use in humans (Berridge 2009; Berridge and Robinson 1998; Berridge et al. 2009; Robinson and Berridge 2000). The phenomenon of behavioral sensitization has been demonstrated for a variety of drugs of addiction, such as amphetamines (Degoulet et al. 2009), cocaine (Burger and Martin-Iverson 1994), and nicotine (Kosowski and Liljequist 2005). It may potentially model elements of drug craving and relapse in humans (Vanderschuren and Kalivas 2000). Although useful for studying several aspects of drug-induced neuroplasticity, the behavioral

sensitization model, like CPP, is limited because animals never experience contingent drug self-administration, a hallmark of human addiction.

The most widely accepted animal model of drug abuse and addiction is the self-administration paradigm. During this operant conditioning procedure, the animal presses a lever (i.e., the operandum), which triggers the delivery of a reward (e.g., cocaine). Animals can be trained to perform a variety of different operant behaviors (e.g., nose pokes) in order to receive the drug after varying amounts of attempts. Like humans, animals will readily make operant responses in order to self-administer most drugs of abuse, including opiates, cannabinoids, alcohol, nicotine, amphetamines, and cocaine (Feltenstein and See 2008; Heidbreder 2011; Yahyavi-Firouz-Abadi and See 2009). Furthermore, studies almost universally demonstrate that animals will preferentially respond on a reinforced (i.e., active), rather than a non-reinforced (i.e., inactive) operandum. This suggests that like humans, animals are able to rapidly discriminate between responses that elicit the delivery of drug and nondrug rewards.

While a variety of species and routes of drug administration can be used, most animal studies of addiction involve the use of rodents or nonhuman primates. Drugs of abuse are typically self-administered intravenously, via a chronic indwelling catheter, or orally. The abuse potential of different compounds in humans is well predicted by animal intravenous self-administration models. This suggests that this paradigm mimics human abuse with greater ecological validity than repeated experimenter-delivered administration (e.g., intraperitoneal, subcutaneous). Therefore, the drug self-administration model appears to possess reasonable face, construct. and predictive validity for examining the neuropharmacological profiles of drugs that are readily abused by humans.

Craving and the recommencement of drug seeking and drug taking following drug abstinence are significant features of addiction (Sinha and Li 2007; Volkow et al. 2002a, 2006b). Factors believed to contribute to drug craving and relapse include exposure to conditioned drug cues, negative mood states, and stress. These triggers have been examined using animal models of relapse that employ an "extinction–reinstatement" approach. In the "extinction–reinstatement model," animals are allowed to self-administer a drug (e.g., cocaine) for prolonged periods of time, mimicking chronic drug use in humans. The animals then undergo extinction training in which the previously reinforced behavior (e.g., pressing a lever) fails to elicit drug delivery. These animals are then exposed to small amounts of the previously administered drug (called drug priming) or environmental stressors (e.g., foot shock) to test the reinstatement of drug self-administration after extinction. Research has shown that conditioned cues, drug priming, and stress are all powerful triggers for the reinstatement of drug-seeking behavior, as indexed by an increase in a behavior previously paired with a drug (Shaham et al. 2003). The reinstatement of drug self-administration is believed to model relapse to drug use in humans. The application of the reinstatement model has also proved useful in examining the neural circuitry underlying drug relapse (Kruzich et al. 2001; Kruzich and See 2001; Weiss et al. 2000).

Neuroimaging Studies in Addiction

Studies Using Positron Emission Tomography (PET)

PET imaging directly assesses neurotransmitter systems in the brain by using a radioactive tracer that recognizes a particular target. There are a number of well characterized tracers for some neurotransmitter systems of interest in addiction (e.g., dopaminergic system), but not for others (e.g., glutamate). This limits the utility of PET investigations.

Dopamine

Cocaine and methamphetamine (or "crystal meth") increase dopamine levels in ways that can be measured by an increase in the displacement of PET tracers that bind to dopamine ($D_{2)}$ receptors (e.g., [^{11}C]raclopride). The increase in dopamine produced by stimulants in healthy volunteers is dose-related and reflects the "high" that people experience (Volkow et al. 1999). These findings, supported by earlier animal studies, showed that drugs of addiction increase dopamine in the NAcc. This led to the view that dopamine release was a necessary, perhaps even sufficient condition, for drugs to have addictive potential. Recent work, however, has cast doubt on this. Heroin and other opioids, nicotine, and cannabis do not appear to produce detectable increases in dopamine (Bossong et al. 2009; Brody et al. 2004; Daglish et al. 2008; Stokes et al. 2009).

In cocaine- and alcohol-dependent individuals, amphetamine- or methylphenidate-stimulated release of dopamine, particularly in the ventral striatum, is blunted compared with healthy volunteers (Martinez et al. 2005). This finding challenges the theory of sensitization, which would predict increased dopamine levels in addicted individuals. Cocaine-dependent individuals also reported a reduced "high" and blunted change in dopamine levels that predicted the choice for cocaine over money (Martinez et al. 2007; Volkow et al. 1997).

The dopamine receptor, however, may play a key role in addiction propensity. Low levels of dopamine receptors are associated with a greater rewarding effect of stimulants (Volkow et al. 1999), while high levels are possibly protective in alcoholism (Volkow et al. 2006a). The use of stimulants (e.g., cocaine and methamphetamine) has been shown to lower dopamine receptor numbers (Dagher et al. 2001; Heinz et al. 2004; Lee et al. 2009; Volkow et al. 2001). Similarly, in alcohol dependence, lower striatal $D_{2/3}$ receptor availability has been reported (Volkow et al. 1996; Martinez et al. 2005). However, there is no evidence of reductions in $D_{2/3}$ receptor availability in cannabis dependence (Sevy et al. 2008) or use (Stokes et al. 2012). A recent study has shown that DAT availability is significantly reduced in the striatum of long-term cannabis users and cigarette smokers (Leroy et al. 2011), suggesting that disturbances in dopamine functioning is associated with chronic use. Daglish et al. 2008 were also unable to detect lower striatal $D_{2/3}$ receptor levels in methadone-maintained, opioid-addicted individuals. Martinez and colleagues did report a reduction in recently abstinent heroin-addicted individuals (Martinez et al. 2011), suggesting that the level of striatal $D_{2/3}$ receptors

may depend on whether opioid dependent individuals are free of or maintained on opioid drugs.

Opioid System

In cocaine addicted individuals, regional brain mOR levels remained elevated in the anterior frontal/cingulate cortex during 12 weeks of abstinence (Gorelick et al. 2005). These regions are strongly implicated in "top-down" cognitive regulation of impulses and behavior. Elevated mOR levels in the medial frontal and middle frontal gyri prior to psychosocial treatment were significantly associated with greater cocaine use during treatment (Ghitza et al. 2010). This study also found that elevations in mOR levels in the anterior cingulate cortex (ACC), medial frontal, and insular cortices correlated with a shorter duration of cocaine abstinence. Significantly, mOR binding was a more powerful predictor of treatment outcome than baseline drug and alcohol use.

The endogenous opioid system plays a significant role in alcohol dependence, as indicated by the efficacy of opiate antagonists (e.g., naltrexone) in pharmacotherapeutic trials (Lingford-Hughes et al. 2012a). Several studies have reported an increased availability of mOR in striatal regions in abstinent alcoholics (Heinz et al. 2005; Williams et al. 2009; Weerts et al. 2011). A similar increase in mOR availability has been found in abstinent opioid-dependent individuals (Williams et al. 2007). An increased availability of mOR has been demonstrated in addiction to a number of pharmacologically different substances of abuse and therefore may be involved in the vulnerability to and perpetuation of drug taking after abstinence.

GABA

Several studies have shown that GABA binding is lower in abstinent alcohol-dependent patients (Abi-Dargham et al. 1998; Lingford-Hughes et al. 1998) or its function reduced (Lingford-Hughes et al. 2005). The reduction in GABA binding may be the result of the downregulation of GABA in order to reduce the impact of sedative drugs on the GABA system. Individuals at risk of alcoholism, and addiction in general, may have preexisting reduced levels of GABA activity. The α_5 subtype of the $GABA_A$ receptor is highly expressed in brain regions that regulate emotion and reward, such as the ventral striatum, and a reduction in these receptors is found in persons with alcoholism (Lingford-Hughes et al. 2012b).

Studies Using Functional MRI (fMRI)

fMRI exploits the fact that the magnetic properties of blood change as oxygen is removed. These changes can be detected using an MRI measure known as the Blood-Oxygen-Level-Dependent (BOLD) signal while a person performs a behavioral task (e.g., reward learning). The BOLD response represents the change in oxyhemoglobin to deoxyhemoglobin ratio in venous blood. The strength of this signal in a brain region (e.g., ventral striatum) indicates the relative level of

oxygenated to deoxygenated blood at that location. Because neuronal activity requires oxygen, the BOLD signal is believed to indirectly reflect neuronal activity at that location during the psychological process being studied.

The use of behavioral assays that specifically tap into the neural circuitry on which drugs of abuse act allows fMRI neuroscientists to explore differences and similarities between the long-term effects of different drugs. The cognitive domains that have been under investigation in recent years, and which have provided some insight into the addicted brain, comprise memory, planning and impulse control, and more subjective experiences such as empathy (see below).

Drug-Related Stimuli

The production of strong emotional and cognitive responses to drug-related stimuli, referred to as cue reactivity, is a common and clinically important feature of drug addiction. Studies have investigated the neural correlates of cue reactivity and craving using cue-exposure techniques, which are ethically less challenging than giving drugs of abuse to addicted individuals. In particular the reactivity of neural reward circuits to drug-related cues has been widely studied to test whether there is an "overvaluation" of drug reinforcers, as has been hypothesized (Goldstein and Volkow 2002). Understanding how this system operates is important because drug-related cues may increase attentional bias and expectancy of drug delivery in both current and abstinent drug users. Such studies may potentially identify neural mechanisms and inform treatment development by providing potential cognitive or pharmacological targets (Muraven 2010; Schoenmakers et al. 2010; Shoptaw et al. 2008; Franklin et al. 2011; Goldstein et al. 2010).

fMRI studies have identified common brain regions (e.g., amygdala, OFC, and VS) that are involved in cue reactivity and craving elicited by drug-related cues in drug-using populations. We describe studies in nicotine addiction to illustrate how such imaging and reactivity has clinical relevance. Greater reactivity to smoking-related images has been reported in the insula and dorsal ACC (dACC) in those nicotine-dependent women who relapsed (Janes et al. 2010). The importance of the insula, which integrates interoceptive (i.e., bodily) states into conscious feelings and decision-making processes involving uncertain risk and reward, has recently emerged with evidence that damage to the insula disrupts nicotine addiction (Naqvi and Bechara 2008). Another study revealed that extinction-based smoking cessation treatment attenuated responses to smoking cues in the amygdala, and the same attenuation pattern in the thalamus predicted which smokers remained abstinent (McClernon et al. 2007). Concerning the impact of medication, varenicline has been shown to reduce responses in the VS and medial OFC to smoking-related cues as well as subjective craving (Franklin et al. 2011).

Imaging with alcohol-related cues has shown activation of similar brain areas. For instance, heavy drinkers show significantly greater activations in the dorsal striatum (DS) than social drinkers, and light drinkers show higher cue-induced activations in the VS and prefrontal areas than heavy social drinkers (Vollstadt-Klein et al. 2010). Detoxified alcoholics have less activation in the VS during the anticipation of nondrug rewards than healthy controls but increased VS activation

in response to alcohol-associated cues (Wrase et al. 2007). This finding suggests that mesolimbic activation in alcoholics (and addiction as a whole) is biased toward the processing of alcohol, as opposed to conventional reward cues, supporting the hypothesis of a reward deficiency syndrome. Various medications such as naltrexone (mOR antagonist), ondansetron (serotonin $5HT_3$ receptor antagonist), and aripiprazole ($D_{2/3}$ partial agonist) all reduce VS activation in response to alcohol-related cues in non-treatment-seeking alcoholics (Myrick et al. 2008, 2010).

In abstinent heroin addicts and cocaine abusers, salient drug-related cues have been shown to result in activation in the ACC in all participants, but PFC activation was only seen in those that experienced craving (Wexler et al. 2001). The study by Daglish et al. (2001) also revealed that ACC activation increased, rather than decreased, with the duration of abstinence. This finding may support the long-held belief that addiction can be an enduring process involving long-term adaptations in various circuits. It has been suggested that increased activity in the OFC reflects a hypersensitivity to reward (Bolla et al. 2003), whereas reduced activity in ACC reflects hyposensitivity to punishment (Garavan and Stout 2005).

As with medications to treat alcoholism, medications used to treat opiate addiction, such as methadone and buprenorphine, have been shown to reduce responses in the insula and hippocampus to salient drug cues (Langleben et al. 2008; Mei et al. 2010). It appears however that activity in the OFC does not dissipate after medication, the impact and clinical implications of which requires further study.

The development of a conditioned, cue-induced neural attentional bias in response to drug-predictive stimuli is accompanied by craving in different drug-using populations. This bias may be implicit in maintaining addictive behaviors and provoking drug relapse among users attempting to remain abstinent. Significantly, functional brain imaging procedures have been shown to reliably measure the effect of neural responses to drug-related stimuli on drug relapse. This may be important when testing the effects of medications on these responses. Future research may improve treatment outcomes in addiction medicine by identifying neural signatures that predict relapse.

Reward Processing

Reward is a central driver of incentive-based learning that elicits appropriate responses to stimuli and shapes the development of goal-directed behaviors. Motivational theories of drug use make different predictions about how drug use may differentially recruit brain areas, such as the VS, in response to rewards (Bjork et al. 2008). The reward deficiency syndrome (RDS) and the allostatic hypotheses (AH), for example, both postulate that addiction is the result of a deficit in dopamine motivational circuitry for nondrug rewards and that only drugs of abuse are able to normalize dopamine at the VS (Blum et al. 2000; Koob et al. 2004). This may induce reflexive, conditioned responses to drug cues and diminish responses to cues that signal nondrug rewards. Alternatively, the “impulsivity hypothesis” of addiction suggests that persons who are vulnerable to, or suffering from, addiction have an excessive approach and reduced inhibitory control over their behavior

(Bechara 2005; Bickel et al. 2007). This hypothesis is supported by longitudinal studies which have shown that both poor self-control and high novelty seeking in childhood are significant predictors of substance use in adolescence (Ding et al. 2004; Masse and Tremblay 1997; Myers et al. 1995) and addiction in later life (Fergusson et al. 2007).

Substance-dependent persons exhibit both impulsive and reward-centered choice behavior, and those with alcohol, cocaine, heroin, and nicotine dependence have an increased preference for small immediate over larger delayed rewards (Bechara et al. 2001; Bickel and Marsch 2001; Bjork et al. 2004; Heil et al. 2006; Reynolds and Fields 2011; Robles et al. 2011). This suggests that individuals who are both prone to, and engage in, chronic substance use have some combination of reward hypersensitivity and deficient inhibitory control (Bechara 2005; Bickel et al. 2007; Solomon and Corbit 1974). Assessing neural responses to nondrug rewards in substance abusers has particular value in evaluating these hypotheses and establishing patterns of reward functioning in addiction.

As described above, significantly lower numbers of D_2 receptors or released dopamine have been found in the striatum of people addicted to alcohol, cocaine, and methamphetamine (Martinez et al. 2005, 2007, 2009, 2011; Volkow et al. 2004). While this evidence is consistent with the RDS and AH, there is less consistent evidence in heroin, nicotine, or opioid addiction. It is not easy to determine whether the impairment precedes or follows addiction. Those at risk of initiating substance abuse may be hyporesponsive to nondrug rewards due to deficient dopamine functioning in the striatum which is overcome by taking drugs of abuse that enhance dopamine levels.

Most fMRI studies in addiction have attempted to examine reward sensitivity using the Monetary Incentive Delay (MID) task (Knutson et al. 2001). The MID task allows researchers to measure brain activation while a person anticipates and receives monetary reward and punishment. The person first views a brief visual cue indicating the type of reward trial they will participate in. This is followed by a short delay after which the person responds to a target stimulus and does or does not receive a reward depending on their response to the target. Significantly, fMRI BOLD responses during the MID delay period correlate with dopamine release in the VS (Schott et al. 2008), appearing to substantiate its sensitivity to dopamine reward functioning.

In alcoholism, support for the RDS/AH is demonstrated by blunted VS responses during reward anticipation compared with nonalcoholics (Beck et al. 2009; Wrase et al. 2007). However, alcoholics did not differ from nonalcoholics during reward anticipation, but they did differ in their responses to reward outcomes (Bjork et al. 2008), a finding more consistent with the impulsivity hypothesis. In cannabis-using populations, there is support for both hypotheses. Greater activation in the VS has been shown in cannabis users than drug-naïve controls during reward anticipation, consistent with the impulsivity hypothesis (Nestor et al. 2010). By contrast, Van Hell and colleagues found that cannabis users had significantly less activation in the VS compared to non-cigarette smokers, but not cigarette smokers (van Hell et al. 2010), thereby supporting the RDS/AH.

Greater activation in the left and right VS, right caudate, and right insula has been shown in treatment-seeking cocaine-addicted individuals using the MID (Jia et al. 2011). Notably, some neural responses predicted treatment success with activation during reward anticipation in the bilateral thalamus and right caudate negatively associated with cocaine-negative urinalyses and activation in the left amygdala and parahippocampal gyrus correlated negatively with treatment retention. These findings suggest that in treatment-seeking cocaine-addicted persons, impulsive corticolimbic reward circuitry for nondrug rewards may be a neural biomarker that predicts treatment outcome (Jia et al. 2011). (Goldstein et al. 2007a) have also reported dysfunctional PFC activation during instrumental tasks in cocaine-addicted persons (Goldstein et al. 2007b). Finally, it has also been shown that in cigarette smokers, neural responses in the VS during reward anticipation are significantly lower than in control subjects (Peters et al. 2011). This has also been observed during the receipt of delayed rewards in this population (Luo et al. 2011).

The concepts of reward and impulsivity are both important in eliciting appropriate responses to stimuli and shaping the development of goal-directed behaviors. Since the empirical findings to date are consistent with both hypotheses, future research on the processing of nondrug rewards in addiction will need to address a potential disparity in neural responses in different types of addictions. In doing so, brain imaging research of reward processing in substance abusers may be able to delineate neural responses that are contingent upon the substance of abuse, the treatment-seeking status of the individual, and the duration of their abstinence.

Cognitive Control and Decision-Making

Flexible goal-directed behavior requires an adaptive cognitive control system for organizing and optimizing processing (Ridderinkhof et al. 2004a, b). Evidence from cognitive neuroscience is beginning to converge on the different contributions of the PFC in cognitive control. This convergence of evidence may identify potential biomarkers of compromised cognition that predict both the initiation and continuation of drug abuse. Since addiction is by definition continued drug use and recurrent drug relapse in the face of serious negative consequences, decrements in cognitive inhibitory control may be a core feature of the disease.

Laboratory tests of cognitive inhibitory control usually involve a person withholding a habitual motor response or ignoring the presentation of irrelevant stimuli while continually updating information and monitoring one's performance. The processes of cognitive inhibitory control and monitoring have consistently been shown to involve the PFC and ACC (Carter et al. 1998; Garavan et al. 1999, 2002; Ullsperger and von Cramon 2001). If the ability to inhibit and monitor one's behavior is important in the development and maintenance of addiction (Garavan and Stout 2005), then brain imaging assessments of cognitive inhibitory control may identify deficits in both behavior and brain functioning.

Dysfunctional activity in the PFC including ACC and OFC of different drug-using (i.e., alcohol, cannabis, cocaine, heroin, methamphetamine and nicotine) individuals has been shown in fMRI studies using tests of cognitive inhibitory control and monitoring when compared with demographically matched drug-naïve

individuals (Volkow et al. 2007; Garavan et al. 2008; Goldstein et al. 2004, 2010; Hester and Garavan 2004; Kaufman et al. 2003). Importantly, it has been shown that severe global cognitive impairment makes cocaine-addicted individuals less amenable to behavioral treatments (Aharonovich et al. 2003, 2006). This underscores the need to uncover biomarkers of cognitive control in addiction to inform rehabilitation programs for individual substance abusers.

Error monitoring has also been shown to be impaired in substance using populations, for example. Reduced functioning in the ACC when cannabis users are required to indicate their awareness of errors (Hester et al. 2009) or in chronic heroin users during error monitoring (Forman et al. 2004). Notably, previous research in early cocaine and methamphetamine abstinence has shown cortical neural deficits during verbal and visuospatial (Kubler et al. 2005) working memory (Moeller et al. 2010; Tomasi et al. 2007), conflict resolution (Nestor et al. 2011a) and decision-making processes (Hoffman et al. 2008; Monterosso et al. 2007). These findings support the notion that disruptions in prefrontal circuits are important for general, flexible, goal-directed behavior. Interestingly one study has shown that ex-smokers who had been abstinent for a year or more had increased lateral PFC activation compared with both smokers and nicotine naïve participants (Nestor et al. 2011b). Increased lateral PFC activation may be an important characteristic of successful abstinence in former smokers.

To assess decision-making in individuals with substance dependence, studies have used the Iowa Gambling Task (IGT) (Bechara et al. 1994). On this task patients with damage to the ventromedial prefrontal cortex (VMPFC) appear to be oblivious to the future consequences of their actions (i.e., myopia for the future), appearing only to be guided by their immediate prospects. Subsequent neuroimaging studies in healthy participants using the IGT have shown increased activation in the VMPFC, ACC, parietal/insular cortices, amygdala, and striatum during the actual decision-making component of the task (Ernst et al. 2002; Matthews et al. 2004; Verney et al. 2003). All of these regions are known to be affected by drugs of abuse. This had led to the “somatic-marker” hypothesis in which decision-making depends on the neural substrates that regulate homeostasis, emotion, and feeling (Verdejo-Garcia and Bechara 2009). According to this model, there should be a link between alterations in processing emotions in substance abusers and their impairments in decision-making.

There is some support for this hypothesis. For example, after 3 weeks abstinence, cocaine abusers have greater activation during performance of the IGT in the right OFC and less activation in the right DLPFC and left medial PFC than a control group (Bolla et al. 2003). These results suggest that cocaine abusers show persistent functional abnormalities in prefrontal neural networks involved in decision-making that may undermine attempts to remain abstinent. In a similar study, after 4 weeks of abstinence cannabis users, particularly heavy users (53–84 joints/week), had greater activation in the left cerebellum and less activation in the right OFC and DLPFC than controls (Bolla et al. 2005). These preliminary findings suggest that prefrontal neural deficits in heavy cannabis users are manifested in decrements in decision-making.

Other studies have examined patterns of regional brain activation in abstinent drug users during decision-making. Imaging patterns in methamphetamine abusing individuals performing the two-choice prediction task have shown decreased activation of the OFC, DLPFC, insular, and inferior parietal cortices (Paulus et al. 2003). These patterns of brain activation were strong predictors of relapse. Here activation patterns in the right insular, posterior cingulate, and temporal cortex obtained in early recovery correctly predicted 90 % of subjects who did not relapse and 94 % of subjects who did (Paulus et al. 2005).

Summary

Our knowledge about how drugs affect brain functioning and neural circuits in abuse and dependence has substantially increased in the last few decades due to developments in neuroimaging. Animal studies have improved in their complexity to better reflect what happens in man. In human and animal studies, the mesolimbic dopaminergic system has continued to receive much attention, and evidence suggests that hypofunctioning in this system is involved in vulnerability to drug liking for naïve users and also to relapse in addicted individuals. The role of other neurotransmitter systems is receiving more attention, and the importance of opioid and GABAergic systems, for instance, is recognized and has led to improvements in clinical treatment. Psychological constructs of behaviors such as reward processing, decision-making, and impulsivity have been widely studied in substance use and abuse with impairments generally described. Challenges for the future include using our knowledge and neuroimaging to bring psychological and pharmacological theories closer together so that the interplay between the impact of the drugs and/or their psychological or pharmacological treatment on underlying psychological processes is clearer.

References

Abi-Dargham, A., Krystal, J. H., Anjilvel, S., Scanley, B. E., Zoghbi, S., Baldwin, R. M., et al. (1998). Alterations of benzodiazepine receptors in type II alcoholic subjects measured with SPECT and [123I]iomazenil. *The American Journal of Psychiatry, 155*(11), 1550–1555.

Aharonovich, E., Nunes, E., & Hasin, D. (2003). Cognitive impairment, retention and abstinence among cocaine abusers in cognitive-behavioral treatment. *Drug and Alcohol Dependence, 71*(2), 207–211.

Aharonovich, E., Hasin, D. S., Brooks, A. C., Liu, X., Bisaga, A., & Nunes, E. V. (2006). Cognitive deficits predict low treatment retention in cocaine dependent patients. *Drug and Alcohol Dependence, 81*(3), 313–322.

Amara, S. G., & Kuhar, M. J. (1993). Neurotransmitter transporters: Recent progress. *Annual Review of Neuroscience, 16*, 73–93.

Bardo, M. T., & Bevins, R. A. (2000). Conditioned place preference: What does it add to our preclinical understanding of drug reward? *Psychopharmacology, 153*(1), 31–43.

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. *Nature Neuroscience, 8*(11), 1458–1463.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1–3), 7–15.

Bechara, A., Dolan, S., Denburg, N., Hindes, A., Anderson, S. W., & Nathan, P. E. (2001). Decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers. *Neuropsychologia, 39*(4), 376–389.

Beck, A., Schlagenhauf, F., Wustenberg, T., Hein, J., Kienast, T., Kahnt, T., et al. (2009). Ventral striatal activation during reward anticipation correlates with impulsivity in alcoholics. *Biological Psychiatry, 66*(8), 734–742.

Benowitz, N. L. (2008). Clinical pharmacology of nicotine: Implications for understanding, preventing, and treating tobacco addiction. *Clinical Pharmacology and Therapeutics, 83*(4), 531–541.

Benowitz, N. L. (2009). Pharmacology of nicotine: Addiction, smoking-induced disease, and therapeutics. *Annual Review of Pharmacology and Toxicology, 49*, 57–71.

Berridge, K. C. (2009). Wanting and liking: Observations from the neuroscience and psychology laboratory. *Inquiry (Oslo), 52*(4), 378.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research. Brain Research Reviews, 28*(3), 309–369.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'Liking', 'wanting', and learning. *Current Opinion in Pharmacology, 9*(1), 65–73.

Bickel, W., & Marsch, L. (2001). Toward a behavioral economic understanding of drug dependence: Delay discounting processes. *Addiction, 96*(1), 73–86.

Bickel, W. K., Miller, M. L., Yi, R., Kowal, B. P., Lindquist, D. M., & Pitcock, J. A. (2007). Behavioral and neuroeconomics of drug addiction: Competing neural systems and temporal discounting processes. *Drug and Alcohol Dependence, 90*(Suppl 1), S85–S91.

Bjork, J. M., Hommer, D. W., Grant, S. J., & Danube, C. (2004). Impulsivity in abstinent alcohol-dependent patients: Relation to control subjects and type 1-/type 2-like traits. *Alcohol, 34*(2–3), 133–150.

Bjork, J. M., Smith, A. R., & Hommer, D. W. (2008). Striatal sensitivity to reward deliveries and omissions in substance dependent patients. *NeuroImage, 42*(4), 1609–1621.

Blum, K., Braverman, E. R., Holder, J. M., Lubar, J. F., Monastra, V. J., Miller, D., et al. (2000). Reward deficiency syndrome: A biogenetic model for the diagnosis and treatment of impulsive, addictive, and compulsive behaviors. *Journal of Psychoactive Drugs, 32*(Suppl i–iv), 1–112.

Bolla, K. I., Eldreth, D. A., London, E. D., Kiehl, K. A., Mouratidis, M., Contoreggi, C., et al. (2003). Orbitofrontal cortex dysfunction in abstinent cocaine abusers performing a decision-making task. *NeuroImage, 19*, 1085–1094.

Bolla, K. I., Eldreth, D. A., Matochik, J. A., & Cadet, J. L. (2005). Neural substrates of faulty decision-making in abstinent marijuana users. *NeuroImage, 26*(2), 480–492.

Bossong, M. G., van Berckel, B. N., Boellaard, R., Zuurman, L., Schuit, R. C., Windhorst, A. D., et al. (2009). Delta 9-tetrahydrocannabinol induces dopamine release in the human striatum. *Neuropsychopharmacology, 34*(3), 759–766.

Brody, A. L., Olmstead, R. E., London, E. D., Farahi, J., Meyer, J. H., Grossman, P., et al. (2004). Smoking-induced ventral striatum dopamine release. *The American Journal of Psychiatry, 161*(7), 1211–1218.

Brody, A. L., Mandelkern, M. A., London, E. D., Olmstead, R. E., Farahi, J., Scheibal, D., et al. (2006). Cigarette smoking saturates brain alpha 4 beta 2 nicotinic acetylcholine receptors. *Archives of General Psychiatry, 63*(8), 907–915.

Bruijnzeel, A. W. (2009). Kappa-Opioid receptor signaling and brain reward function. *Brain Research Reviews, 62*(1), 127–146.

Burger, L. Y., & Martin-Iverson, M. T. (1994). Increased occupation of D1 and D2 dopamine receptors accompanies cocaine-induced behavioral sensitization. *Brain Research, 639*(2), 228–232.

Cami, J., & Farre, M. (2003). Drug addiction. *The New England Journal of Medicine, 349*(10), 975–986.

Carter, C., Braver, T., Barch, D., Botvinick, M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*(5364), 747–749.

Curran, H. V., Brignell, C., Fletcher, S., Middleton, P., & Henry, J. (2002). Cognitive and subjective dose-response effects of acute oral Delta 9-tetrahydrocannabinol (THC) in infrequent cannabis users. *Psychopharmacology, 164*(1), 61–70.

D'Souza, D. C., Braley, G., Blaise, R., Vendetti, M., Oliver, S., Pittman, B., et al. (2008). Effects of haloperidol on the behavioral, subjective, cognitive, motor, and neuroendocrine effects of Delta-9-tetrahydrocannabinol in humans. *Psychopharmacology, 198*(4), 587–603.

Dagher, A., Bleicher, C., Aston, J. A., Gunn, R. N., Clarke, P. B., & Cumming, P. (2001). Reduced dopamine D1 receptor binding in the ventral striatum of cigarette smokers. *Synapse, 42*(1), 48–53.

Daglish, M. R., Weinstein, A., Malizia, A. L., Wilson, S., Melichar, J. K., Britten, S., et al. (2001). Changes in regional cerebral blood flow elicited by craving memories in abstinent opiate-dependent subjects. *The American Journal of Psychiatry, 158*(10), 1680–1686.

Daglish, M. R. C., Williams, T., Wilson, S. J., Taylor, L. G., Brooks, D. J., Myles, J. S., Grasby, P. G., Lingford-Hughes, A. R., & Nutt, D. J. (2008). No measurable dopamine response to heroin in the brains of human addicts. *The British Journal of Psychiatry, 193*(1), 65–72.

Degoulet, M. F., Rostain, J. C., David, H. N., & Abraini, J. H. (2009). Repeated administration of amphetamine induces a shift of the prefrontal cortex and basolateral amygdala motor function. *The International Journal of Neuropsychopharmacology, 12*(7), 965–974.

Ding, Y. S., Gatley, S. J., Thanos, P. K., Shea, C., Garza, V., Xu, Y., et al. (2004). Brain kinetics of methylphenidate (Ritalin) enantiomers after oral administration. *Synapse, 53*(3), 168–175.

Drenan, R. M., Grady, S. R., Steele, A. D., McKinney, S., Patzlaff, N. E., McIntosh, J. M., et al. (2010). Cholinergic modulation of locomotion and striatal dopamine release is mediated by alpha6alpha4* nicotinic acetylcholine receptors. *Journal of Neuroscience, 30*(29), 9877–9889.

Ernst, M., Bolla, K., Mouratidis, M., Contoreggi, C., Matochik, J. A., Kurian, V., et al. (2002). Decision-making in a risk-taking task: A PET study. *Neuropsychopharmacology, 26*(5), 682–691.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience, 8*(11), 1481–1489.

Feltenstein, M. W., & See, R. E. (2008). The neurocircuitry of addiction: An overview. *British Journal of Pharmacology, 154*(2), 261–274.

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2007). Conduct and attentional problems in childhood and adolescence and later substance use, abuse and dependence: Results of a 25-year longitudinal study. *Drug and Alcohol Dependence, 88*(Suppl 1), S14–S26.

Forman, S. D., Dougherty, G. G., Casey, B. J., Siegle, G. J., Braver, T. S., Barch, D. M., et al. (2004). Opiate addicts lack error-dependent activation of rostral anterior cingulate. *Biological Psychiatry, 55*(5), 531–537.

Franklin, T., Wang, Z., Suh, J. J., Hazan, R., Cruz, J., Li, Y., et al. (2011). Effects of varenicline on smoking cue-triggered neural and craving responses. *Archives of General Psychiatry, 68*(5), 516–526.

Garavan, H., & Stout, J. C. (2005). Neurocognitive insights into substance abuse. *Trends in Cognitive Science, 9*(4), 195–201.

Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America, 96*(14), 8301–8306.

Garavan, H., Ross, T. J., Murphy, K., Roche, R. A., & Stein, E. A. (2002). Dissociable executive functions in the dynamic control of behavior: Inhibition, error detection, and correction. *NeuroImage, 17*(4), 1820–1829.

Garavan, H., Kaufman, J. N., & Hester, R. (2008). Acute effects of cocaine on the neurobiology of cognitive control. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1507), 3267–3276.
Ghitza, U. E., Preston, K. L., Epstein, D. H., Kuwabara, H., Endres, C. J., Bencherif, B., et al. (2010). Brain mu-opioid receptor binding predicts treatment outcome in cocaine-abusing outpatients. *Biological Psychiatry, 68*(8), 697–703.
Goldstein, R. Z., & Volkow, N. D. (2002). Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *The American Journal of Psychiatry, 159*, 1642–1652.
Goldstein, R. Z., Leskovjan, A. C., Hoff, A. L., Hitzemann, R., Bashan, F., Khalsa, S. S., et al. (2004). Severity of neuropsychological impairment in cocaine and alcohol addiction: Association with metabolism in the prefrontal cortex. *Neuropsychologia, 42*(11), 1447–1458.
Goldstein, R. Z., Tomasi, D., Alia-Klein, N., Cottone, L. A., Zhang, L., Telang, F., et al. (2007a). Subjective sensitivity to monetary gradients is associated with frontolimbic activation to reward in cocaine abusers. *Drug and Alcohol Dependence, 87*(2–3), 233–240.
Goldstein, R. Z., Tomasi, D., Rajaram, S., Cottone, L. A., Zhang, L., Maloney, T., et al. (2007b). Role of the anterior cingulate and medial orbitofrontal cortex in processing drug cues in cocaine addiction. *Neuroscience, 144*(4), 1153–1159.
Goldstein, R. Z., Woicik, P. A., Maloney, T., Tomasi, D., Alia-Klein, N., Shan, J., et al. (2010). Oral methylphenidate normalizes cingulate activity in cocaine addiction during a salient cognitive task. *Proceedings of the National Academy of Sciences of the United States of America, 107*(38), 16667–16672.
Gonzales, D., Rennard, S. I., Nides, M., Oncken, C., Azoulay, S., Billing, C. B., et al. (2006). Varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs sustained-release bupropion and placebo for smoking cessation: A randomized controlled trial. *JAMA: The Journal of the American Medical Association, 296*(1), 47–55.
Gorelick, D. A., Kim, Y. K., Bencherif, B., Boyd, S. J., Nelson, R., Copersino, M., et al. (2005). Imaging brain mu-opioid receptors in abstinent cocaine users: Time course and relation to cocaine craving. *Biological Psychiatry, 57*(12), 1573–1582.
Grady, S. R., Salminen, O., McIntosh, J. M., Marks, M. J., & Collins, A. C. (2010). Mouse striatal dopamine nerve terminals express alpha4alpha5beta2 and two stoichiometric forms of alpha4beta2*-nicotinic acetylcholine receptors. *Journal of Molecular Neuroscience, 40*(1–2), 91–95.
Gray, R., Rajan, A. S., Radcliffe, K. A., Yakehiro, M., & Dani, J. A. (1996). Hippocampal synaptic transmission enhanced by low concentrations of nicotine. *Nature, 383*(6602), 713–716.
Heidbreder, C. (2011). Advances in animal models of drug addiction. *Current Topics in Behavioral Neurosciences, 7*, 213–250.
Heil, S. H., Johnson, M. W., Higgins, S. T., & Bickel, W. K. (2006). Delay discounting in currently using and currently abstinent cocaine-dependent outpatients and non-drug-using matched controls. *Addictive Behaviors, 31*(7), 1290–1294.
Heimer, L., & Van Hoesen, G. W. (2006). The limbic lobe and its output channels: Implications for emotional functions and adaptive behavior. *Neuroscience and Biobehavioral Reviews, 30*(2), 126–147.
Heinz, A., Siessmeier, T., Wrase, J., Hermann, D., Klein, S., Grusser, S. M., et al. (2004). Correlation between dopamine D(2) receptors in the ventral striatum and central processing of alcohol cues and craving. *The American Journal of Psychiatry, 161*(10), 1783–1789.
Heinz, A., Reimold, M., Wrase, J., Hermann, D., Croissant, B., Mundle, G., et al. (2005). Correlation of stable elevations in striatal mu-opioid receptor availability in detoxified alcoholic patients with alcohol craving: A positron emission tomography study using carbon 11-labeled carfentanil. *Archives of General Psychiatry, 62*(1), 57–64.
Heishman, S. J., Schuh, K. J., Schuster, C. R., Henningfield, J. E., & Goldberg, S. R. (2000). Reinforcing and subjective effects of morphine in human opioid abusers: Effect of dose and alternative reinforcer. *Psychopharmacology, 148*(3), 272–280.

Herkenham, M., Lynn, A. B., Johnson, M. R., Melvin, L. S., de Costa, B. R., & Rice, K. C. (1991). Characterization and localization of cannabinoid receptors in rat brain: A quantitative in vitro autoradiographic study. *Journal of Neuroscience, 11*(2), 563–583.

Hester, R., & Garavan, H. (2004). Executive dysfunction in cocaine addiction: Evidence for discordant frontal, cingulate, and cerebellar activity. *Journal of Neuroscience, 24*(49), 11017–11022.

Hester, R., Nestor, L., & Garavan, H. (2009). Impaired error awareness and anterior cingulate cortex hypoactivity in chronic cannabis users. *Neuropsychopharmacology, 34*(11), 2450–2458.

Hill, J. L., & Zacny, J. P. (2000). Comparing the subjective, psychomotor, and physiological effects of intravenous hydromorphone and morphine in healthy volunteers. *Psychopharmacology, 152*(1), 31–39.

Hoffman, W. F., Schwartz, D. L., Huckans, M. S., McFarland, B. H., Meiri, G., Stevens, A. A., et al. (2008). Cortical activation during delay discounting in abstinent methamphetamine dependent individuals. *Psychopharmacology, 201*(2), 183–193.

Howell, L. L., & Kimmel, H. L. (2008). Monoamine transporters and psychostimulant addiction. *Biochemical Pharmacology, 75*(1), 196–217.

Huestis, M. A., Boyd, S. J., Heishman, S. J., Preston, K. L., Bonnet, D. L., Fur, G., et al. (2007). Single and multiple doses of rimonabant antagonize acute effects of smoked cannabis in male cannabis users. *Psychopharmacology, 194*(4), 505–515.

Hunault, C. C., Mensinga, T. T., de Vries, I., Kelholt-Dijkman, H. H., Hoek, J., Kruidenier, M., et al. (2008). Delta-9-tetrahydrocannabinol (THC) serum concentrations and pharmacological effects in males after smoking a combination of tobacco and cannabis containing up to 69 mg THC. *Psychopharmacology, 201*(2), 171–181.

Janes, A. C., Pizzagalli, D. A., Richardt, S., deB Frederick, B., Chuzi, S., Pachas, G., et al. (2010). Brain reactivity to smoking cues prior to smoking cessation predicts ability to maintain tobacco abstinence. *Biological Psychiatry, 67*(8), 722–729.

Jia, Z., Worhunsky, P. D., Carroll, K. M., Rounsaville, B. J., Stevens, M. C., Pearlson, G. D., et al. (2011). An initial study of neural responses to monetary incentives as related to treatment outcome in cocaine dependence. *Biological Psychiatry, 70*(6), 553–560.

Jorenby, D. E., Hays, J. T., Rigotti, N. A., Azoulay, S., Watsky, E. J., Williams, K. E., et al. (2006). Efficacy of varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs placebo or sustained-release bupropion for smoking cessation: A randomized controlled trial. *JAMA: The Journal of the American Medical Association, 296*(1), 56–63.

Kaufman, J. N., Ross, T. J., Stein, E. A., & Garavan, H. (2003). Cingulate hypoactivity in cocaine users during a GO-NOGO task as revealed by event-related functional magnetic resonance imaging. *Journal of Neuroscience, 23*(21), 7839–7843.

Klink, R., de Kerchove d'Exaerde, A., Zoli, M., & Changeux, J. P. (2001). Molecular and physiological diversity of nicotinic acetylcholine receptors in the midbrain dopaminergic nuclei. *Journal of Neuroscience, 21*(5), 1452–1463.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience, 21*(16), RC159.

Koob, G. F. (2004). A role for GABA mechanisms in the motivational effects of alcohol. *Biochemical Pharmacology, 68*(8), 1515–1525.

Koob, G. F., & Le Moal, M. (1997). Drug abuse: Hedonic homeostatic dysregulation. *Science, 278*(5335), 52–58.

Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology, 35*(1), 217–238.

Koob, G. F., Ahmed, S. H., Boutrel, B., Chen, S. A., Kenny, P. J., Markou, A., et al. (2004). Neurobiological mechanisms in the transition from drug use to drug dependence. *Neuroscience and Biobehavioral Reviews, 27*(8), 739–749.

Kosowski, A. R., & Liljequist, S. (2005). Behavioural sensitization to nicotine precedes the onset of nicotine-conditioned locomotor stimulation. *Behavioural Brain Research, 156*(1), 11–17.

Kruzich, P. J., & See, R. E. (2001). Differential contributions of the basolateral and central amygdala in the acquisition and expression of conditioned relapse to cocaine-seeking behavior. *Journal of Neuroscience, 21*(14), RC155.

Kruzich, P. J., Congleton, K. M., & See, R. E. (2001). Conditioned reinstatement of drug-seeking behavior with a discrete compound stimulus classically conditioned with intravenous cocaine. *Behavioral Neuroscience, 115*(5), 1086–1092.

Kubler, A., Murphy, K., & Garavan, H. (2005). Cocaine dependence and attention switching within and between verbal and visuospatial working memory. *European Journal of Neuroscience, 21*(7), 1984–1992.

Kuhar, M. J., Ritz, M. C., & Boja, J. W. (1991). The dopamine hypothesis of the reinforcing properties of cocaine. *Trends in Neurosciences, 14*(7), 299–302.

Langleben, D. D., Ruparel, K., Elman, I., Busch-Winokur, S., Pratiwadi, R., Loughead, J., et al. (2008). Acute effect of methadone maintenance dose on brain FMRI response to heroin-related cues. *The American Journal of Psychiatry, 165*(3), 390–394.

Lee, B., London, E. D., Poldrack, R. A., Farahi, J., Nacca, A., Monterosso, J. R., et al. (2009). Striatal dopamine d2/d3 receptor availability is reduced in methamphetamine dependence and is linked to impulsivity. *Journal of Neuroscience, 29*(47), 14734–14740.

Leroy, C., Karila, L., Martinot, J. L., Lukasiewicz, M., Duchesnay, E., Comtat, C., et al. (2011). Striatal and extrastriatal dopamine transporter in cannabis and tobacco addiction: A high-resolution PET study. *Addiction Biology*, *17*(6):981–90.

Lingford-Hughes, A. R., Acton, P. D., Gacinovic, S., Suckling, J., Busatto, G. F., Boddington, S. J. A., Bullmore, E., Woodruff, P. W., Costa, D. C., Pilowsky, L. S., Ell, P. J., Marshall, E. J., & Kerwin, R. W. (1998). Reduced levels of the GABA-benzodiazepine receptor in alcohol dependency in the absence of grey matter atrophy. *The British Journal of Psychiatry, 173*, 116–122.

Lingford-Hughes, A. R., Wilson, S. J., Cunningham, V. J., Feeney, A., Stevenson, B., Brooks, D. J., & Nutt, D. J. (2005). GABA-benzodiazepine receptor function in alcohol dependence: A combined 11C-flumazenil PET and pharmacodynamic study. *Psychopharmacology, 180*, 595–606.

Lingford-Hughes, A. R., Watson, B., Kalk, N., & Reid, A. (2010). Neuropharmacology of addiction and how it informs treatment. *British Medical Bulletin, 96*, 93–110.

Lingford-Hughes, A., Welch, S., Peters, L., Nutt, D. on behalf of expert group. (2012a). Evidence-based guidelines for the pharmacological management of substance misuse, addiction and comorbidity: Recommendations from BAP. *Journal of Psychopharmacology, 26*(7), 899–952.

Lingford-Hughes, A. R., Reid, A. G., Myers, J., Feeney, A., Hammers, A., Taylor, L. G., Rosso, L., Turkheimer, F., Brooks, D. J., Grasby, P., & Nutt, D. J. (2012b). A [^{11}C]Ro15 4513 PET study suggests that alcohol dependence in man is associated with reduced a5 benzodiazepine receptors in limbic regions. *Journal of Psychopharmacology, 26*(2), 273–281.

Lippiello, P., Letchworth, S. R., Gatto, G. J., Traina, V. M., & Bencherif, M. (2006). Ispronicline: A novel alpha4beta2 nicotinic acetylcholine receptor-selective agonist with cognition-enhancing and neuroprotective properties. *Journal of Molecular Neuroscience, 30*(1–2), 19–20.

Luo, S., Ainslie, G., Giragosian, L., & Monterosso, J. R. (2011). Striatal hyposensitivity to delayed rewards among cigarette smokers. *Drug and Alcohol Dependence, 116*(1–3), 18–23.

Mameli-Engvall, M., Evrard, A., Pons, S., Maskos, U., Svensson, T. H., Changeux, J. P., et al. (2006). Hierarchical control of dopamine neuron-firing patterns by nicotinic receptors. *Neuron, 50*(6), 911–921.

Mansvelder, H. D., & McGehee, D. S. (2002). Cellular and synaptic mechanisms of nicotine addiction. *Journal of Neurobiology, 53*(4), 606–617.

Martinez, D., Gil, R., Slifstein, M., Hwang, D. R., Huang, Y., Perez, A., et al. (2005). Alcohol dependence is associated with blunted dopamine transmission in the ventral striatum. *Biological Psychiatry, 58*(10), 779–786.

Martinez, D., Narendran, R., Foltin, R. W., Slifstein, M., Hwang, D. R., Broft, A., et al. (2007). Amphetamine-induced dopamine release: Markedly blunted in cocaine dependence and predictive of the choice to self-administer cocaine. *The American Journal of Psychiatry, 164*(4), 622–629.

Martinez, D., Greene, K., Broft, A., Kumar, D., Liu, F., Narendran, R., et al. (2009). Lower level of endogenous dopamine in patients with cocaine dependence: Findings from PET imaging of D (2)/D(3) receptors following acute dopamine depletion. *The American Journal of Psychiatry, 166*(10), 1170–1177.

Martinez, D., Saccone, P. A., Liu, F., Slifstein, M., Orlowska, D., Grassetti, A., et al. (2011). Deficits in dopamine D(2) receptors and presynaptic dopamine in heroin dependence: Commonalities and differences with other types of addiction. *Biological Psychiatry, 71*(3):192–8.

Masse, L. C., & Tremblay, R. E. (1997). Behavior of boys in kindergarten and the onset of substance use during adolescence. *Archives of General Psychiatry, 54*(1), 62–68.

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine, 3*(11), e442.

Matthews, S. C., Simmons, A. N., Lane, S. D., & Paulus, M. P. (2004). Selective activation of the nucleus accumbens during risk-taking decision making. *Neuroreport, 15*(13), 2123–2127.

McClernon, F. J., Hiott, F. B., Liu, J., Salley, A. N., Behm, F. M., & Rose, J. E. (2007). Selectively reduced responses to smoking cues in amygdala following extinction-based smoking cessation: Results of a preliminary functional magnetic resonance imaging study. *Addiction Biology, 12*(3–4), 503–512.

Mei, W., Zhang, J. X., & Xiao, Z. (2010). Acute effects of sublingual buprenorphine on brain responses to heroin-related cues in early-abstinent heroin addicts: An uncontrolled trial. *Neuroscience, 170*(3), 808–815.

Moeller, F. G., Steinberg, J. L., Schmitz, J. M., Ma, L., Liu, S., Kjome, K. L., et al. (2010). Working memory fMRI activation in cocaine-dependent subjects: Association with treatment response. *Psychiatry Research, 181*(3), 174–182.

Monterosso, J. R., Ainslie, G., Xu, J., Cordova, X., Domier, C. P., & London, E. D. (2007). Frontoparietal cortical activity of methamphetamine-dependent and comparison subjects performing a delay discounting task. *Human Brain Mapping, 28*(5), 383–393.

Muraven, M. (2010). Practicing self-control lowers the risk of smoking lapse. *Psychology of Addictive Behaviors, 24*(3), 446–452.

Myers, M. G., Brown, S. A., & Mott, M. A. (1995). Preadolescent conduct disorder behaviors predict relapse and progression of addiction for adolescent alcohol and drug abusers. *Alcoholism, Clinical and Experimental Research, 19*(6), 1528–1536.

Myrick, H., Anton, R. F., Li, X., Henderson, S., Randall, P. K., & Voronin, K. (2008). Effect of naltrexone and ondansetron on alcohol cue-induced activation of the ventral striatum in alcohol-dependent people. *Archives of General Psychiatry, 65*(4), 466–475.

Myrick, H., Li, X., Randall, P. K., Henderson, S., Voronin, K., & Anton, R. F. (2010). The effect of aripiprazole on cue-induced brain activation and drinking parameters in alcoholics. *Journal of Clinical Psychopharmacology, 30*(4), 365–372.

Naqvi, N. H., & Bechara, A. (2008). The hidden island of addiction: The insula. *Trends in Neurosciences*, *32*(1):56–67.

Nestor, L., Hester, R., & Garavan, H. (2010). Increased ventral striatal BOLD activity during non-drug reward anticipation in cannabis users. *NeuroImage, 49*(1), 1133–1143.

Nestor, L., McCabe, E., Jones, J., Clancy, L., & Garavan, H. (2011a). Differences in "bottom-up" and "top-down" neural activity in current and former cigarette smokers: Evidence for neural substrates which may promote nicotine abstinence through increased cognitive control. *NeuroImage, 56*(4), 2258–2275.

Nestor, L. J., Ghahremani, D. G., Monterosso, J., & London, E. D. (2011b). Prefrontal hypoactivation during cognitive control in early abstinent methamphetamine-dependent subjects. *Psychiatry Research, 194*(3), 287–295.

Nutt, D. J., & Lingford-Hughes, A. R. (2008). Addiction: The clinical interface. *British Journal of Pharmacology, 154*(2), 397–405.
Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology, 47*(6), 419–427.
Pastor, V., Andres, M. E., & Bernabeu, R. O. (2012). The effect of previous exposure to nicotine on nicotine place preference. *Psychopharmacology (Berlin)*, *226*(3):551–60.
Patterson, F., Jepson, C., Strasser, A. A., Loughead, J., Perkins, K. A., Gur, R. C., et al. (2009). Varenicline improves mood and cognition during smoking abstinence. *Biological Psychiatry, 65*(2), 144–149.
Paulus, M. P., Hozack, N., Frank, L., Brown, G. G., & Schuckit, M. A. (2003). Decision making by methamphetamine-dependent subjects is associated with error-rate-independent decrease in prefrontal and parietal activation. *Biological Psychiatry, 53*(1), 65–74.
Paulus, M. P., Tapert, S. F., & Schuckit, M. A. (2005). Neural activation patterns of methamphetamine-dependent subjects during decision making predict relapse. *Archives of General Psychiatry, 62*(7), 761–768.
Peters, J., Bromberg, U., Schneider, S., Brassen, S., Menz, M., Banaschewski, T., et al. (2011). Lower ventral striatal activation during reward anticipation in adolescent smokers. *The American Journal of Psychiatry, 168*(5), 540–549.
Peto, R., Lopez, A. D., Boreham, J., Thun, M., Heath, C., Jr., & Doll, R. (1996). Mortality from smoking worldwide. *British Medical Bulletin, 52*(1), 12–21.
Reynolds, B., & Fields, S. (2011). Delay discounting by adolescents experimenting with cigarette smoking. *Addiction*, *107*(2):417–24.
Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004a). The role of the medial frontal cortex in cognitive control. *Science, 306*(5695), 443–447.
Ridderinkhof, K. R., van den Wildenberg, W. P., Segalowitz, S. J., & Carter, C. S. (2004b). Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition, 56*(2), 129–140.
Robinson, T. E., & Berridge, K. C. (2000). The psychology and neurobiology of addiction: An incentive-sensitization view. *Addiction, 95*(Suppl 2), S91–S117.
Robles, E., Huang, B. E., Simpson, P. M., & McMillan, D. E. (2011). Delay discounting, impulsiveness, and addiction severity in opioid-dependent patients. *Journal of Substance Abuse Treatment, 41*(4), 354–362.
Rudnick, G., & Clark, J. (1993). From synapse to vesicle: The reuptake and storage of biogenic amine neurotransmitters. *Biochimica et Biophysica Acta, 1144*(3), 249–263.
Schoenmakers, T. M., de Bruin, M., Lux, I. F., Goertz, A. G., Van Kerkhof, D. H., & Wiers, R. W. (2010). Clinical effectiveness of attentional bias modification training in abstinent alcoholic patients. *Drug and Alcohol Dependence, 109*(1–3), 30–36.
Schott, B. H., Minuzzi, L., Krebs, R. M., Elmenhorst, D., Lang, M., Winz, O. H., et al. (2008). Mesolimbic functional magnetic resonance imaging activations during reward anticipation correlate with reward-related ventral striatal dopamine release. *Journal of Neuroscience, 28*(52), 14311–14319.
Sevy, S., Smith, G. S., Ma, Y., Dhawan, V., Chaly, T., Kingsley, P. B., et al. (2008). Cerebral glucose metabolism and D(2)/D (3) receptor availability in young adults with cannabis dependence measured with positron emission tomography. *Psychopharmacology, 197*(4), 549–556.
Shaham, Y., Shalev, U., Lu, L., De Wit, H., & Stewart, J. (2003). The reinstatement model of drug relapse: History, methodology and major findings. *Psychopharmacology, 168*(1–2), 3–20.
Shoptaw, S., Heinzerling, K. G., Rotheram-Fuller, E., Kao, U. H., Wang, P. C., Bholat, M. A., et al. (2008). Bupropion hydrochloride versus placebo, in combination with cognitive behavioral therapy, for the treatment of cocaine abuse/dependence. *Journal of Addictive Diseases, 27*(1), 13–23.

Sieghart, W. (2006). Structure, pharmacology, and function of GABAA receptor subtypes. *Advances in Pharmacology, 54*, 231–263.

Sinha, R., & Li, C. S. (2007). Imaging stress- and cue-induced drug and alcohol craving: Association with relapse and clinical implications. *Drug and Alcohol Review, 26*(1), 25–31.

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation. I. Temporal dynamics of affect. *Psychological Review, 81*(2), 119–145.

Sora, I., Li, B., Fumushima, S., Fukui, A., Arime, Y., Kasahara, Y., et al. (2009). Monoamine transporter as a target molecule for psychostimulants. *International Review of Neurobiology, 85*, 29–33.

Sticht, M., Mitsubata, J., Tucci, M., & Leri, F. (2010). Reacquisition of heroin and cocaine place preference involves a memory consolidation process sensitive to systemic and intra-ventral tegmental area naloxone. *Neurobiology of Learning and Memory, 93*(2), 248–260.

Stokes, P. R., Mehta, M. A., Curran, H. V., Breen, G., & Grasby, P. M. (2009). Can recreational doses of THC produce significant dopamine release in the human striatum? *NeuroImage, 48*(1), 186–190.

Stokes, P. R. A., Egerton, A., Watson, B., Reid, A., Lappin, J., Nutt, D., & Lingford-Hughes, A. (2012). History of cannabis use is not associated with alterations in striatal dopamine D2/D3 receptor availability. *Journal of Psychopharmacology, 26*(1), 144–149.

Sullivan, J. M., Risacher, S. L., Normandin, M. D., Yoder, K. K., Froehlich, J. C., & Morris, E. D. (2011). Imaging of alcohol-induced dopamine release in rats: Preliminary findings with [(11) C]raclopride PET. *Synapse, 65*(9), 929–937.

Tang, A., George, M. A., Randall, J. A., & Gonzales, R. A. (2003). Ethanol increases extracellular dopamine concentration in the ventral striatum in C57BL/6 mice. *Alcoholism, Clinical and Experimental Research, 27*(7), 1083–1089.

Thorn, D. A., Winter, J. C., & Li, J. X. (2012). Agmatine attenuates methamphetamine-induced conditioned place preference in rats. *European Journal of Pharmacology, 680*(1–3), 69–72.

Tomasi, D., Goldstein, R. Z., Telang, F., Maloney, T., Alia-Klein, N., Caparelli, E. C., et al. (2007). Widespread disruption in brain activation patterns to a working memory task during cocaine abstinence. *Brain Research, 1171*, 83–92.

Tsou, K., Brown, S., Sanudo-Pena, M. C., Mackie, K., & Walker, J. M. (1998). Immunohistochemical distribution of cannabinoid CB1 receptors in the rat central nervous system. *Neuroscience, 83*(2), 393–411.

Ullsperger, M., & von Cramon, D. Y. (2001). Subprocesses of performance monitoring: A dissociation of error processing and response competition revealed by event-related fMRI and ERPs. *NeuroImage, 14*(6), 1387–1401.

van Hell, H. H., Vink, M., Ossewaarde, L., Jager, G., Kahn, R. S., & Ramsey, N. F. (2010). Chronic effects of cannabis use on the human reward system: An fMRI study. *European Neuropsychopharmacology, 20*(3), 153–163.

Vanderschuren, L. J., & Kalivas, P. W. (2000). Alterations in dopaminergic and glutamatergic transmission in the induction and expression of behavioral sensitization: A critical review of preclinical studies. *Psychopharmacology, 151*(2–3), 99–120.

Verdejo-Garcia, A., & Bechara, A. (2009). A somatic marker theory of addiction. *Neuropharmacology, 56*(Suppl 1), 48–62.

Verney, S. P., Brown, G. G., Frank, L., & Paulus, M. P. (2003). Error-rate-related caudate and parietal cortex activation during decision making. *Neuroreport, 14*(7), 923–928.

Volkow, N. D., Wang, G. J., Fowler, J. S., Logan, J., Hitzemann, R., Ding, Y. S., et al. (1996). Decreases in dopamine receptors but not in dopamine transporters in alcoholics. *Alcoholism, Clinical and Experimental Research, 20*(9), 1594–1598.

Volkow, N. D., Wang, G. J., Fowler, J. S., Logan, J., Gatley, S. J., Hitzemann, R., et al. (1997). Decreased striatal dopaminergic responsiveness in detoxified cocaine-dependent subjects. *Nature, 386*(6627), 830–833.

Volkow, N. D., Wang, G. J., Fowler, J. S., Logan, J., Gatley, S. J., Wong, C., et al. (1999). Reinforcing effects of psychostimulants in humans are associated with increases in brain dopamine and occupancy of D(2) receptors. *Journal of Pharmacology and Experimental Therapeutics, 291*(1), 409–415.

Volkow, N. D., Chang, L., Wang, G. J., Fowler, J. S., Ding, Y. S., Sedler, M., et al. (2001). Low level of brain dopamine D2 receptors in methamphetamine abusers: Association with metabolism in the orbitofrontal cortex. *The American Journal of Psychiatry, 158*(12), 2015–2021.

Volkow, N., Fowler, J., Wang, G., & Goldstein, R. (2002a). Role of dopamine, the frontal cortex and memory circuits in drug addiction: Insight from imaging studies. *Neurobiology of Learning and Memory, 78*(3), 610–624.

Volkow, N. D., Wang, G. J., Maynard, L., Fowler, J. S., Jayne, B., Telang, F., et al. (2002b). Effects of alcohol detoxification on dopamine D2 receptors in alcoholics: A preliminary study. *Psychiatry Research, 116*(3), 163–172.

Volkow, N. D., Fowler, J. S., Wang, G. J., & Swanson, J. M. (2004). Dopamine in drug abuse and addiction: Results from imaging studies and treatment implications. *Molecular Psychiatry, 9*(6), 557–569.

Volkow, N. D., Wang, G. J., Begleiter, H., Porjesz, B., Fowler, J. S., Telang, F., et al. (2006a). High levels of dopamine D2 receptors in unaffected members of alcoholic families: Possible protective factors. *Archives of General Psychiatry, 63*(9), 999–1008.

Volkow, N. D., Wang, G. J., Telang, F., Fowler, J. S., Logan, J., Childress, A. R., et al. (2006b). Cocaine cues and dopamine in dorsal striatum: Mechanism of craving in cocaine addiction. *Journal of Neuroscience, 26*(24), 6583–6588.

Volkow, N. D., Wang, G. J., Telang, F., Fowler, J. S., Logan, J., Jayne, M., et al. (2007). Profound decreases in dopamine release in striatum in detoxified alcoholics: Possible orbitofrontal involvement. *Journal of Neuroscience, 27*(46), 12700–12706.

Vollstadt-Klein, S., Wichert, S., Rabinstein, J., Buhler, M., Klein, O., Ende, G., et al. (2010). Initial, habitual and compulsive alcohol use is characterized by a shift of cue processing from ventral to dorsal striatum. *Addiction, 105*(10), 1741–1749.

Weerts, E. M., Wand, G. S., Kuwabara, H., Munro, C. A., Dannals, R. F., Hilton, J., et al. (2011). Positron emission tomography imaging of mu- and delta-opioid receptor binding in alcohol-dependent and healthy control subjects. *Alcoholism, Clinical and Experimental Research, 35*(12), 2162–2173.

Weiss, F., Maldonado-Vlaar, C. S., Parsons, L. H., Kerr, T. M., Smith, D. L., & Ben-Shahar, O. (2000). Control of cocaine-seeking behavior by drug-associated stimuli in rats: Effects on recovery of extinguished operant-responding and extracellular dopamine levels in amygdala and nucleus accumbens. *Proceedings of the National Academy of Sciences of the United States of America, 97*(8), 4321–4326.

Wexler, B. E., Gottschalk, C. H., Fulbright, R. K., Prohovnik, I., Lacadie, C. M., Rounsaville, B. J., et al. (2001). Functional magnetic resonance imaging of cocaine craving. *The American Journal of Psychiatry, 158*(1), 86–95.

Williams, T. M., Daglish, M. R. C., Lingford-Hughes, A. R., Taylor, L. G., Hammers, A., Brooks, D. J., Grasby, P. G., Myles, J. S., & Nutt, D. J. (2007). Increased availability of opioid receptors in early abstinence from opioid dependence: A [11C]diprenorphine PET study. *The British Journal of Psychiatry, 191*(1), 63–69.

Williams, T. M., Davies, S. J., Taylor, L. G., Daglish, M. R., Hammers, A., Brooks, D. J., Nutt, D. J., & Lingford-Hughes, A. (2009). Brain opioid receptor binding in early abstinence from alcohol dependence and relationship to craving: An [(11)C]diprenorphine PET study. *European Neuropsychopharmacology, 19*(10), 740–748.

Wise, R. A. (1996). Neurobiology of addiction. *Current Opinion in Neurobiology, 6*(2), 243–251.

Wise, R. A. (2002). Brain reward circuitry: Insights from unsensed incentives. *Neuron, 36*(2), 229–240.

Wise, R. A. (2005). Forebrain substrates of reward and motivation. *The Journal of Comparative Neurology, 493*(1), 115–121.

Wooltorton, J. R., Pidoplichko, V. I., Broide, R. S., & Dani, J. A. (2003). Differential desensitization and distribution of nicotinic acetylcholine receptor subtypes in midbrain dopamine areas. *Journal of Neuroscience, 23*(8), 3176–3185.

Woolverton, W. L., & Johnson, K. M. (1992). Neurobiology of cocaine abuse. *Trends in Pharmacological Sciences, 13*(5), 193–200.

Wrase, J., Schlagenhauf, F., Kienast, T., Wustenberg, T., Bermpohl, F., Kahnt, T., et al. (2007). Dysfunction of reward processing correlates with alcohol craving in detoxified alcoholics. *NeuroImage, 35*(2), 787–794.

Yahyavi-Firouz-Abadi, N., & See, R. E. (2009). Anti-relapse medications: Preclinical models for drug addiction treatment. *Pharmacology & Therapeutics, 124*(2), 235–247.

Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response

66

Wayne D. Hall, Adrian Carter, and Murat Yücel

Contents

W.D. Hall (✉)
The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia

Queensland Brain Institute, The University of Queensland, St Lucia, QLD, Australia
e-mail: w.hall@uq.edu.au

A. Carter
The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia
e-mail: adrian.carter@uq.edu.au

M. Yücel
Monash Clinical and Imaging Neuroscience, School of Psychology and Psychiatry, Monash University, Melbourne, VIC, Australia
e-mail: murat.yucel@monash.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_69,

Abstract

Brain imaging research in addiction promises to provide neurobiological and functional markers that may improve treatment of alcoholism and other types of drug addiction. More speculatively, it may also enable us to prevent addiction by intervening early with individuals who are identified as being at increased risk. In this chapter, we review the research findings on the use of neuroimaging to identify those at greater risk of developing addiction and to match addicted individuals to the treatments that are most likely to assist them toward abstinence. We then discuss the ethical and public policy issues that may arise from the clinical use of these technologies. We consider issues such as: (i) the commercialization of neuroimaging via direct-to-consumer marketing before the technology has been properly validated; (ii) the misuse of neuroimaging to discriminate against individuals at increased risk of developing addiction; (iii) the possible benefits and risks of "medicalizing" drug use and addiction, including possible effects on stigmatization of, and discrimination against, drug-dependent persons; and (iv) the possible misuse of neurobiological theories of addiction by those marketing alcohol and tobacco to undermine public health strategies that aim to reduce the population-level harms these substances cause. Finally, we examine arguments that evaluations of the future predictive utility of neuroimaging in the field of addiction will require substantial investments in health services research to evaluate the cost-effectiveness of this approach. Ethical assessments of the proposed applications of neuroimaging research should be an integral part of this health services research.

Introduction

Over the past several decades, research into the neural correlates of addiction has shown that specific brain circuits are differentially activated by addictive drugs and drug-related cues and that these patterns of activation differ between persons who are, and are not, addicted to drugs (Ersche and Robbins 2011; Parvaz et al. 2011). See ► Chap. 65, "Neuroscience Perspectives on Addiction: Overview" in this section. These findings raise the possibility that imaging brain anatomy and functional circuitry may have two types of predictive applications in the field of addiction: (1) predicting how addicted individuals will respond to different types of treatments for their addiction (Ho et al. 2010; Hutchison 2010; Reske and Paulus 2011; Singh and Rose 2009) and, more speculatively, (2) identifying deviant brain characteristics in childhood or adolescence that predict an individual's susceptibility to developing addiction, potentially enabling addiction to be prevented (e.g., Duka et al. 2011; Ersche et al. 2011; Loth et al. 2011; Schumann et al. 2010). In this chapter, we critically discuss the likelihood of these applications being realized. We also consider the social and ethical issues that may arise if these forms of prediction – which we term *neuroprediction* for short – do prove possible.

We first describe the type of neuroimaging biomarkers that are most likely to be used to predict addiction susceptibility or treatment outcomes. We then briefly review what may be gleaned from research on the predictive value of genomic information (Evans et al. 2011; Gartner et al. 2009; Hall et al. 2010) as many of the issues in genetic prediction and neuroprediction of disease risk are likely to be similar (Evans et al. 2011; Hall et al. 2010; Ioannidis et al. 2010). We then consider some of the issues that may arise if these potential applications of neuroimaging are realized and made commercially available before the evidence is available to support their validity. Many of these issues have also arisen in debates about the risks and benefits of medicalizing human behavior, an approach that is facilitated by the provision of neurobiological explanations of disapproved forms of human behavior such as drug use and addiction (Campbell 2012; Midanik 2006; Netherland 2011).

Potential Applications of Neuroprediction in Addiction

The most immediate applications of neuroimaging are likely to be the clinical matching of addiction treatment to specific addicted persons to optimize outcomes (Reske and Paulus 2011). Recent findings from longitudinal neuroimaging studies in dementia provide a possible model for the clinical use of imaging in addiction. The rate and pattern of change in brain size and connections between brain regions can better distinguish between patients with dementia and control participants than simple differences in these measures at baseline (Pievani et al. 2011). These patterns appear to reflect underlying pathological changes in brain tissue that can distinguish between persons with different subtypes of dementia and also identify non-symptomatic cases of dementia to allow for earlier intervention (Pievani et al. 2011; Seeley et al. 2009). In the case of addiction, functional neuroimaging might predict which individuals require additional treatment (e.g., depot medication, a drug vaccine, additional social support, targeted neurocognitive training) to reduce relapse (Gu et al. 2010; Sinha 2011; Sutherland et al. 2012). These methods may also be used in correctional settings (see ▶ Chap. 107, “Prediction of Antisocial Behavior”) to predict the likelihood that paroled prisoners will return to drug use after their release (Nadelhoffer et al. 2010). In this chapter, we concentrate on potential uses of neuroprediction in voluntary treatment settings to avoid the additional ethical challenges introduced by their use within the criminal justice system.

The use of neuroimaging to predict future addiction risk in young adults is a much more speculative possibility. The research required to enable such prediction (e.g., large-scale prospective and longitudinal follow-up studies spanning 10–20 years) has not been completed, although it is beginning (e.g., Cheetham et al. 2012; Schumann et al. 2010). The aims of this research are to: (i) identify persons at increased risk of developing addiction and (ii) provide preventive or early intervention programs to reduce this risk. Assuming that it proves feasible to do these things, “predictive neuroimaging” will raise many more ethical issues than neuroimaging for treatment matching. Specifically, prediction will need

to be accurate enough to minimize the harms to persons who are mistakenly identified as being at risk of developing a socially stigmatized disorder (false positives).

Risk prediction is ethically justified if there is an effective intervention that reduces the likelihood of a disorder, with minimal adverse consequences, in those at high risk (Ioannidis 2009). The expert consensus is that we do not have effective medical or behavioral interventions to prevent addiction in high-risk persons (Babor et al. 2010a; Carter and Hall 2012). In the absence of such interventions, risk prediction would expose high-risk individuals to potential stigmatization and discrimination without any therapeutic benefits that might justify this risk. Experience with genetic prediction of disease risk suggests that simply telling a person that he or she is at increased risk is unlikely to change their behavior in ways that will reduce the risk (e.g., by avoiding drug use in the case of addiction) (Marteau et al. 2010). It is also not clear whether risk information based on brain imaging will prove more persuasive in reducing risky behavior than simpler information that is already available to persons at risk, namely, a family history of alcohol or drug problems.

We assume that any future use of neuroprediction in the field of addiction will use actuarial, that is, the mechanical combination of statistical information about the risk for future outcomes often used by insurance and financial institutions, rather than relying solely on the clinical judgments made by medical experts. This is because the preponderance of evidence suggests that clinical judgments perform uniformly poorer at this type of prediction than actuarial methods (Dawes et al. 1989). In making these predictions, neuroimaging data is likely to be combined with personal history and clinical information in a standardized way (e.g., using statistical multivariate pattern recognition or "classification" algorithms) rather than the unstandardized and unaided clinical judgment of an individual clinician (Klöppel et al. 2012; Nadelhoffer et al. 2010). The practical utility of this approach will depend on demonstrating that there is a greater improvement in prediction when neuroimaging information is combined with other information rather than when it is used as a single predictor. Health services research will be required to decide whether neuroimaging improves on simpler methods of prediction and if it does, whether the size of the improvement is sufficient to justify the additional costs of using neuroimaging.

Cautionary Experiences from Genomics

Some neuroscientists have expressed optimism about the future utility of neuroimaging in predicting the risk profile and treatment outcomes for addiction and other psychiatric disorders (Loth et al. 2011; Schumann et al. 2010). For instance, they have suggested that functional brain imaging of children or adolescents while they perform certain cognitive tasks will identify those at increased risk of developing addiction if they use drugs during adolescence (e.g., because they have poorly functioning inhibitory control circuits and/or a highly responsive

reward system) (Volkow and Li 2005). It may well prove possible to use imaging biomarkers to make predictions about addiction risk at a *group level* but experiences with genomic prediction of disease risk suggest much more caution when making predictions about the future utility of neuroprediction *in individual cases*.

There were similar high hopes for the predictive use of genomic information in medicine at the time that human genome project was nearing completion (Collins 1999). In the case of addiction, the realization of this hope seemed justified by twin and adoption studies which suggested that genetic factors make a substantial contribution to addiction susceptibility (Agrawal et al. 2012; Bierut 2011). The major challenge for genomic prediction of addiction risk, and disease risk more generally (Hall et al. 2010), has been the failure to identify specific mutations that strongly predict individual addiction risk (Agrawal et al. 2012; Bierut 2011). Large-scale Genome Wide Association Studies (GWAS) and meta-analyses of GWAS have generally identified large numbers of alleles that only weakly predict disease risk (Agrawal et al. 2012; Evans et al. 2011; Swendsen and Le Moal 2011). These alleles also appear to predict an increased risk of developing one or more of a cluster of correlated externalizing behaviors, such as antisocial behavior, drug use, precocious sexual activity, and aggressive acts to others, rather than predicting the risks of specific types of addiction (Edwards et al. 2009). Even when information from risk alleles is actuarially combined, genetic prediction of addiction risks often does not improve on prediction using crude information on family history of addiction (Gartner et al. 2009). This outcome is not surprising given the ontological gap between nuclear DNA and the behavioral phenotype, the likelihood that environmental experiences change gene expression (epigenetics), the probable role played by gene–environment correlations, and the possible role of gene–gene and gene–environment interactions in addiction (Agrawal et al. 2012).

Genetic information seems more useful in predicting treatment response in addicted persons and thereby allowing individuals to be matched to the treatment that is most likely to be effective for them (e.g., Chen et al. 2012). There is some evidence, for example, that common alleles differentially predict treatment outcome in persons with nicotine, alcohol, and opioid dependence (Agrawal et al. 2012). If these results are replicated, health services research will be required to compare the cost-effectiveness of treatment matching using genetic data with simpler treatment matching strategies. The latter could include, for example, giving all patients the most effective treatment first and reserving more intensive treatment for those who fail at first-line treatments.

Challenges in Neuroprediction of Addiction Risk

Neuroprediction of addiction risk and treatment response will probably share some of the challenges faced by genomic risk prediction. It will also face additional challenges that arise from the cost and impracticality of routinely using neuroimaging.

First, neuroimaging research studies are at risk of repeating the experience of publication bias that affected early studies in disease genomics. "Promising results" found in small sample studies were more likely to be published than failures to find differences, raising hopes about prediction that were not borne out in subsequent attempts to replicate in studies using larger samples (Ioannidis 2012). For example, a meta-analysis of studies of brain volume abnormalities in persons with a variety of psychiatric disorders found that more statistically significant results were reported than should have been the case given the small study samples and the small average differences between cases and controls (Ioannidis 2011). This suggests that the literature on this topic has been biased by the selective publication of false positive results in small studies that overestimate effect sizes. More generally, meta-analyses of studies of potential biomarkers in medicine have typically found that estimates of effect sizes for consistently replicated findings *decrease* as the number of published studies increases (Ioannidis 2013).

Second, the current state of neuroimaging research resembles that of genomic research in the first years after completion of the first draft of the human genome. There has been an exponential growth in the generation of large volumes of neuroimaging data in the absence of any consensus among investigators on how to standardize neuroimaging methods, cognitive tasks, and methods of data analysis. This has led to the publication of findings that are often perplexing and difficult to understand. It will be some time before a consensus emerges on these issues that will enable the sort of progress that has occurred in genomics after the advent of GWAS using very large pooled samples from multiple case–control studies, standardized genomic methods, and appropriate statistical methods to address serious multiple comparison problems.

Third, neuroimaging does have a major disadvantage compared to genotyping. Large-scale genotyping studies are relatively inexpensive and easier to conduct than the large-scale longitudinal neuroimaging studies that are required to test the predictive utility of neurobiological markers of addiction risk (e.g., the IMAGEN studies of Whelan et al. 2012). The cost of performing functional magnetic resonance imaging (fMRI) and positron emission topography (PET) scans are also unlikely to decline as quickly as the costs of genome scans have done. Neuroimaging requires individuals to attend specialist imaging facilities to undergo neuroimaging. The neuroimaging procedure is much more complicated, time consuming, and costly than it is to take the blood or saliva samples that enable large numbers of subjects to be genotyped cheaply using fast-throughput, next-generation genomic sequencing. Neuroimaging often measures changes in brain activity in response to behavioral or cognitive tasks that can be difficult to administer in uncooperative participants.

Fourth, it seems most likely that neuroimaging will have utility in clinical settings when matching addicted individuals to customized treatments (Reske and Paulus 2011). For instance, Paulus and colleagues (2005) showed that brain activation patterns obtained using fMRI were able to predict relapse in persons with methamphetamine dependence. Schutz (2008) has also reported that neuroimaging findings predict relapse risk in smokers. If these results are

replicated, we will need to see whether they also predict treatment outcome in other forms of drug dependence.

Fifth, the cost and logistical challenges of neuroimaging will be a barrier to its routine use in addiction treatment matching, even if matching proves possible in research studies. Large-scale neuroimaging will be practicable in tertiary research and treatment settings but primary care physicians will not be able to use these technologies or receive imaging results in a timely way from specialist imaging facilities. It seems more likely that treatment matching will be done using neuropsychological or behavioral tests that neuroimaging studies have shown to measure functioning in key brain regions (e.g., those that are associated with impulsive behavior and relapse to drug use or other compulsive behaviors) (e.g., Fontenelle et al. 2011; Goudriaan et al. 2008). This could include, for example, tests of attentional bias toward drug cues, impulsivity, and the inability to delay reward. These tests would be less expensive, and if they had equal or similar efficacy, would be more cost-effective than neuroimaging in predicting treatment outcome (Cox et al. 2002; Goudriaan et al. 2008; Paulus et al. 2005). Health services research will be required to establish the clinical and scientific validity of neurocognitive tests before they could be routinely used in clinical practice.

In order to assess the clinical utility of neuroimaging in addiction treatment matching, evidence will be needed to show that: (1) these methods predict different responses to various treatments for drug dependence (e.g., different drug treatments or psychological therapies) and (2) matching patients to treatments using this information is more cost-effective than giving everyone the treatment that is the most effective on average (i.e., regardless of genotype or neuroimaging results) (Hall 2007). In order to be cost-effective, these instruments will need to reliably predict differential treatment responses that are large enough to justify the additional costs of testing.

These evaluations will also need to assess the effects that this information has on addicted persons' beliefs about their ability to abstain from drug use. Studies suggest that smokers interpret genetic risk information to mean that they can only quit using biological interventions, if they are able to quit at all (Cappella et al. 2005; Wright et al. 2003). In these studies, smokers who accepted that genetic factors contributed to their cigarette smoking were less confident about quitting and more likely to believe that they needed a biological intervention to quit. Researchers will also need to assess whether neuroimaging information discourages quit attempts in those who try and fail to quit. Patients arguably have a right to be told about genetic risks information but providing this needs to be done in a way so as to avoid discouraging further quitting attempts because most addicted persons only achieve lasting abstinence after a number of attempts to quit (John et al. 2004).

Prediction of Addiction Risk

We do not have research evidence on whether neuroimaging in adolescence can predict addiction risk in young adulthood and we are unlikely to have such information

for some time. For the moment this use of neuroprediction is speculative. Nonetheless, there are some research findings that suggest that this may prove possible.

Longitudinal studies suggest that it is possible to predict the risk of addictive disorders in adulthood from characteristics measured in childhood and adolescence. Moffitt et al. (2011), for example, found that indicators of poor self-control in early childhood and adolescence predicted an increased risk of substance use disorders and poorer health, lower wealth, and an increased risk of police arrest in adulthood. Moffitt and colleagues measured impulsivity, conscientiousness, self-regulation, delay of gratification, inattention, hyperactivity, executive function, will power, and inter-temporal choice starting at the age of 3 and throughout childhood and adolescence in a birth cohort of 1,000 New Zealanders. Their study had a retention rate of over 90 % and there was very little selective attrition of those with lower self-control. The study also collected rich data on study participants and their families from infancy to adulthood. Participant self-control was assessed by measures of impulsive behavior (e.g., impulsive aggression, lack of persistence, and hyperactivity) that were collected at 3, 5, 7, 9, and 11 years of age. These included: researcher observations of behavior at age 3, teacher and parent ratings of behavior during childhood, and participants' self-reports of their behavior during adolescence and early adulthood. The study assessed the relationships between these measures and indicators of health, wealth, and criminal history at the age of 32.

Poor self-control that persisted throughout childhood and adolescence predicted poorer physical and mental health, lower wealth, and a higher risk of having been convicted of a criminal offence by age 32. Alcohol, tobacco, and cannabis use disorders occurred more often among those who were more impulsive in childhood and adolescence. On all these outcomes, the risks of poorer adult outcomes increased steeply as indicators of self-control declined. These relationships persisted after controlling for IQ and higher socioeconomic status, both of which were correlated with poorer self-control in childhood and with poorer adult outcomes.

These findings were supported by more detailed analyses of the data. Study participants whose self-control improved between childhood and young adulthood had better outcomes at age 32 than participants whose self-control did not improve in adulthood. Similarly, individuals with poor self-control in childhood who avoided adolescent "snares" (e.g., starting to smoke cigarettes, leaving school early, and adolescent pregnancy) did better than peers with poor self-control who did not avoid these snares. In fact, even behavioral ratings of self-control made by researchers at age 3 predicted poorer adult outcomes at age 32, although this prediction was poorer than in young people who showed persistent poor self-control throughout childhood and adolescence.

Neuroimaging research has suggested a possible neurobiological basis for the differences in impulsivity found by Moffitt et al. Ersche et al. (2012), for example, reported similar deficits in the neural tracts involved in impulsivity in non-substance-using individuals and their stimulant-dependent sibling. The brain scans of both sibling groups differed from a control group of unrelated healthy individuals who were matched for age and intelligence. The sibling pairs also showed larger deficits in

behavioral regulation than controls on the stop-signal task, a commonly used measure of inhibitory control. The authors argued that stimulant-abusing individuals have structural differences in the striatum and frontal cortices, brain regions that are associated with impaired behavioral control, making them more likely to abuse stimulant drugs. The fact that siblings with the same differences did not abuse stimulants indicates that other factors were also important. A correlation between the extent of impaired brain structure and the duration of stimulant drug use suggested that chronic stimulant abuse can produce these changes in brain structure.

A recent fMRI study has also found an association between self-reported adolescent drug use and patterns of activation in several cortical and subcortical neural networks involved in the inhibition of behavior, including regions implicated in impulsivity (e.g., the basal ganglia, pre-supplementary motor area, orbitofrontal cortices) (Whelan et al. 2012). In this study, the brains of 1,896 adolescents were imaged using fMRI while they performed a stop-signal task to identify the main brain regions involved in inhibitory control. Activation of this inhibitory network was correlated with self-reported alcohol, tobacco, and illicit drug use. This indicated that hypofunctioning in an orbitofrontal network was associated with the initiation of drug use, while right inferior frontal activity was related to the speed of the inhibition process and the use of illegal drugs.

These early results raise the possibility that neuroprediction of addiction risk may one day be possible. More authoritative assessments of the feasibility of neuroprediction require larger, prospective studies like that of the IMAGEN study that is currently following a cohort of 2,000 European adolescents. Participants have already undergone functional and structural neuroimaging studies in early adolescence and will be followed into adulthood. This type of study is required to decide whether differences in brain structure and function are causes or consequences of addiction. If brain function differs in adolescence between those who do and do not develop an addiction, then further studies will need to assess: whether these differences are large enough to reliably predict individual addiction risk; whether neuroprediction improves upon prediction using behavioral and self-report data, such as that collected in the study by Moffitt and colleagues (2011); and whether any improvements in prediction is sufficient to justify the cost of neuroimaging. Evaluations of the performance of neuroprediction will also need to attend to ethical issues that are raised by the prediction of a stigmatized form of behavior such as addiction (Garnett et al. 2011).

Ethical and Public Policy Implications

Premature Commercialization: Direct-to-Consumer Neuroimaging

Putative neuropredictive tests are already being marketed directly to consumers in private psychiatric treatment programs in the USA (Farah 2012). These practices parallel the direct to consumer (DTC) marketing of genetic tests that claim to predict disease risk even though these tests have not been clinically

validated (Mathews et al. 2012). In the USA, for example, the Amen Clinics (http://www.amenclinics.com/) have advertised directly to consumers that their neuroimaging provides novel diagnoses that predict the medical treatments to which individuals are most likely to respond. These clinics claim that tens of thousands of patients have undergone this type of neuroimaging (at their own expense) but present no evidence from the peer-reviewed literature to support the therapeutic claims made. The success of these forms of DTC neuroimaging will depend in large part on public misunderstandings of the information that neuroimaging provides.

Challenges for Public Understanding

Media representations of the role played by genetics in disease may be deterministic. While this may not be true for all media reports (Condit et al. 1998), some media reports claim that research has identified "the gene for addiction" (BBC News 2004; Doyle 2004), with the implication that people with this gene are very likely to develop addiction, and that those who do not have the gene are at low risk of doing so (Khoury et al. 2000). Public understandings of neuroimaging have been similarly simplified (see ▶ Chap. 92, "Neuroscience, Neuroethics, and the Media"), with newspaper articles claiming that neuroscientists have identified the "brain's addiction centre" (BBC News 2007). Misunderstandings of neuroimaging are arguably more likely due to: the novelty and complexity of the technology, the mysterious nature of the physical and biological processes that are measured, the close relationship between brain and behavior, and common misunderstandings of what brain images depict. Media stories on neuroimaging include multi-colored images of brain regions that are said to "light up" when people engage in a cognitive task or respond to drug-related stimuli (Racine et al. 2005, 2006). These are presented as if they were photographs of brain activity that enabled the content of individuals' brains to be directly read (Racine et al. 2005, 2006).

The journalists, and the public who read their articles, are unaware of the large amount of data processing that is required to produce these color-coded brain images (Illes and Racine 2005; Roskies 2008; Schleim and Roiser 2009). For example, it is not well understood that fMRI does not measure brain activation directly, but rather changes in blood oxygenation that occur a small time after brain areas have been activated. Nor is it understood that these changes reflect *relative* differences in activation (rather than absolute levels of brain activation) between different brain regions while a person performs two or more tasks (e.g., an experimental and a control task) (Bell and Racine 2009).

Neuroimages are produced by sophisticated statistical processing that requires numerous assumptions. The patterns of blood oxygenation produced are the product of statistical algorithms that highlight differences between regions in activation. The familiar multi-hued brain images are then produced by fitting *averages* of individual brain scans onto a standardized brain atlas and using arbitrary colors to indicate differences in level of activation between areas of the brain. The inferential

gap between brain activity and brain image is therefore much larger than that between a conventional photograph and its subject (Logothetis 2008; Schleim and Roiser 2009).

It is unclear to what extent the public perceives brain images as fixed, biological features of individuals. Neuroimaging research suggests a more plastic view in which brain images (whether functional or structural) can change fairly rapidly, contradicting the understandings of the brain as an organ with a limited capacity to regenerate and rewire. For example, numerous studies have demonstrated that exercise triggers the production of new neural cells (neurogenesis) (van Praag 2009). Brain imaging studies of London taxi drivers, who are required to memorize the large number of London street names and routes, showed that they develop much larger hippocampi – the regions of the brain responsible for memorizing factual knowledge – than matched controls (Maguire et al. 2000). Neuroimaging studies have also identified learning-related structural changes in the brain in: medical students studying for exams (Draganski et al. 2006), individuals experienced in meditating (Lazar et al. 2005), and persons learning new behaviors (Zatorre et al. 2012). This research suggests a more optimistic view about brain malleability and the potential for recovery from addiction, particularly if researchers can discern how to enhance these changes. This more optimistic view is now being communicated to the public in popular books about the "plastic brain" (e.g., Arrowsmith-Young 2012; Doidge 2007). The popularity of "brain training" games that claim to use neuroplasticity to prevent age-related cognitive decline suggests that the public may be receptive to this view.

We need to do more research on public understanding of neuroimaging and its policy implications. Some critics have argued that the public is overly persuaded by neuroimaging, and neuroscience research more generally (Nadelhoffer et al. 2010; Sinnott-Armstrong et al. 2008), in attaching too much significance to brain images and drawing erroneous inferences when presented with spurious neuroimaging results (Burge 2010; Quart 2012). More recent studies of the impact of neuroimages on mock juries suggest that the public may be more responsive to skeptical views of neuroimaging when such evidence is subject to cross-examination in mock courts (Schweitzer et al. 2011). A recent study has highlighted that it is not just the public that are swayed by neuroscientific explanations. Evidence presented to 180 US judges supporting a neurobiological explanation of a hypothetical psychopathic criminal act significantly reduced the severity of sentences (Aspinwall et al. 2012). It remains to be discovered whether better public education about neuroimaging technologies, and more informed media coverage, will reduce common misinterpretations and generate a more informed public, political representatives, and policy makers. This is an important area in need of good social and psychological research.

Benefits and Risks of Medicalizing Addictive Behavior

Neuroprediction fits well with the medicalization of human behavior, that is, a tendency to explain problematic human behavior in neurobiological terms and to

treat persons displaying these behaviors using drugs to intervene in their putatively disordered brain chemistry (Conrad 2007). Neurobiological explanations of addiction readily fit in with the idea that biological interventions, such as pharmacotherapies, are needed to treat addicted persons and possibly to prevent individuals who are at risk of addiction from developing it.

Any benefits to individuals that may flow from improved prediction of addiction risk presuppose that there are effective forms of prevention or early intervention available to reduce the likelihood that high-risk individuals will develop addiction. We do not have effective preventive interventions at this stage. Interventions such as educating young people about the harms of drug use have, at best, modest effects on their drug use, and not always for the better (Babor et al. 2010b). In the absence of effective preventive interventions, there is little benefit in identifying those who are at increased risk of developing addiction. In the absence of effective prevention, the neuroprediction of addiction risk raises many of the ethical and social concerns expressed by critics of the medicalization of other behavioral disorders (Rose 2006; Rose 2010).

Social scientists, sociologists, and ethicists (e.g., Ashcroft et al. 2007; Buchman et al. 2011; Choudhury et al. 2009; Netherland 2011; Press 2006; Verweij 1999) have argued that medicalization overemphasizes the biological origins of addictive (and other) socially disapproved behavior. Moreover, they argue, it does so at the expense of social and psychological explanations and it acts in ways that may adversely affect people who engage in these behaviors, or are perceived to be at risk of doing so (Caron et al. 2005). If addictions are seen as genetic and neurobiological disorders, for example, these critics argue more resources will be devoted to medical interventions and less attention to social policies, such as imposing higher taxes, banning promotions, and restricting access for young adults (Carlsten and Burke 2006; Chapman and MacKenzie 2010; Evans et al. 2011; Hall and Chikritzhs 2011; Merikangas and Risch 2003).

Critics also argue that neurobiological explanations may adversely affect addicted persons by, for example, reducing their confidence in their ability to quit smoking, drinking, or drug taking (Backlar 1996; Caron et al. 2005; Chapman and MacKenzie 2010). These critics also fear that increased acceptance of a neurobiological view will further stigmatize alcohol- and nicotine-dependent persons (Buchman et al. 2011; Caron et al. 2005). According to these critics, neuroimaging risk information could, like genetic risk information, lead to institutionalized discrimination by employers, who may decline to employ persons at increased addiction risk, and health and life insurers, who decline to cover those identified as being at increased risk of developing disease (Anderlik and Rothstein 2001; Geppert and Roberts 2005; Greely 2001; Hall and Rich 2000; Rothenberg et al. 1997). There is an extensive literature on these issues in the use of genetic prediction of disease risk (Heinrichs 2012; Rothstein and Anderlik 2001), but research is needed to assess to what extent ethical concerns about genetic testing also apply to the neuroprediction of addiction risk.

An argument has been made by Nadelhoffer et al. (2010) for using neuroprediction to improve prediction of re-offending in the criminal justice system. They argue that officials in the criminal justice system already make predictions about individuals, such as sentencing convicted offenders and deciding whether to release prisoners who have served part of their sentences. Moreover, many of these predictions are made using clinical judgment which we know is a very poor way of making decisions. Better predictions can be made using actuarial methods to combine information that includes neuroimaging data. They acknowledge the concerns about third-party uses of neuroimaging data but argue that these concerns are not peculiar to neuroprediction; the same issues arise whenever we attempt to predict risk.

A priority for future research on the utility of neuroprediction should be assessing the extent and the seriousness of the adverse effects of the medicalization that have been raised by social scientists. We need to know, for example, more about the severity of existing forms of stigmatization and their social effects on persons with addictive disorders. We also need to assess any adverse effects of stigma on persons who are identified as being at increased risk of addictive disorders (e.g., among parents or siblings). And we need to know more about the impact that neurobiological explanations of addiction have on addicted persons' interest in quitting and in their confidence in their ability to quit, should they make the attempt.

Subversive Policy Uses of Biological Risk Information

Critics of medicalization argue that neurobiological explanations of addictive behavior may replace effective population-based alcohol and tobacco strategies with high-risk strategies of intervention (Carlsten and Burke 2006; Merikangas and Risch 2003; Miller et al. 2012; Willett 2002). Population-based tobacco control strategies include taxing cigarettes and reducing opportunities to smoke, policies that have halved rates of smoking in Australia (White et al. 2003) and the USA (Pierce et al. 1998) over the past three decades. At present, it makes more policy sense to reduce cigarette smoking using these strategies than it does to spend scarce resources on identifying those at higher risk of becoming nicotine dependent or developing tobacco-related diseases, if they smoke tobacco (Hall et al. 2002; Khoury et al. 2004; Vos et al. 2010).

Public health professionals are concerned about the misuse of biological concepts of addiction risk by industries that want to promote tobacco use and harmful forms of alcohol use (Gundle et al. 2010; Miller et al. 2012). These industries have a history of using individual risk information to undermine public health policies that reduce the use of their products (Hall et al. 2008). Tobacco industry documents (Gundle et al. 2010), for example, show that the industry funded genetic research on smoking and tobacco-related disease in the 1970s and 1980s with the aim of locating the risks of smoking in the genome of the smoker

rather than in tobacco smoking (Gundle et al. 2010). The alcohol industry has also promoted the idea that alcohol-related problems are confined to a minority of genetically vulnerable drinkers (Hall 2005) and that these problems are best addressed by intervening with problem drinkers rather than public health measures that increase alcohol taxes and reduce the availability of alcohol (Babor et al. 2010b). The gambling industry has funded research into the neurobiology of problem gambling (Vrecko 2008), presumably for similar reasons, namely, to locate responsibility for the problem with affected individuals and shift policy attention away from regulating gambling (Schüll 2006).

None of these criticisms justify a ban on neuroimaging or other neurobiological research on any form of addiction, but caution against uncritically accepting the type of public policy implications that powerful economic interests will want to use this research to promote, namely, to locate the problem of addiction solely within the brains of affected individuals rather than modifying the type of social environments that promote the use of addictive commodities.

Conclusions

The failure to realize optimistic predictions about the genomic prediction of addiction risk should caution against similarly optimistic projections about the utility of neuroimaging in predicting addiction risk. Even if these methods prove able to predict individual addiction risk, the costs of screening large numbers of individuals in order to identify the small number at high risk is difficult to justify in the absence of effective preventive interventions. Population health strategies such as increased taxation and reduced opportunities to smoke or drink alcohol are likely to remain more efficient preventive strategies in reducing addiction to nicotine and alcohol.

Any future predictive use of neuroimaging information on addiction risk will need to address community concerns about privacy and the third-party use of such risk information. Public education is also needed about neuroimaging to reduce common misinterpretations of its findings. Research will also be needed on how best to present neuroimaging in ways that do not undermine successful public health strategies for reducing addiction and disease risk (McBride et al. 2010).

The most likely practical benefits of brain imaging in addiction will be in clinical settings. There, imaging may identify homogenous subtypes of addictive disorders that respond preferentially to different types of treatment. Evaluations of the utility of these applications of neuroimaging will require substantial investments in research to identify predictors of addiction treatment response. Assuming that these can be identified, health services research will then be required to evaluate their utility and cost-effectiveness. Such research should include analyses of the ethical issues that may be raised by future uses of neuroimaging.

Acknowledgments We would like to thank Ben Capps, Jayne Lucke, and Eric Racine for helpful comments on an earlier draft of this chapter.

Cross-References

▶ Neuroscience, Neuroethics, and the Media
▶ Neuroscience Perspectives on Addiction: Overview
▶ Prediction of Antisocial Behavior

References

Agrawal, A., Verweij, K. J. H., Gillespie, N. A., Heath, A. C., Lessov-Schlaggar, C. N., Martin, N. G., Nelson, E. C., Slutske, W. S., Whitfield, J. B., & Lynskey, M. T. (2012). The genetics of addiction – a translational perspective. *Translational Psychiatry, 2*, e140.

Anderlik, M. R., & Rothstein, M. A. (2001). Privacy and confidentiality of genetic information: What rules for the new science? *Annual Review of Genomics and Human Genetics, 2*, 401–433.

Arrowsmith-Young, B. (2012). *The woman who changed her brain: And other inspiring stories of pioneering brain transformation*. New York: Free Press.

Ashcroft, R., Campbell, A., & Capps, B. (2007). Ethical aspects of developments in neuroscience and drug addiction. In D. Nutt, T. Robbins, G. Stimson, et al. (Eds.), *Drugs and the future: Brain science, addiction and society* (pp. 439–466). London: Academic.

Aspinwall, L. G., Brown, T. R., & Tabery, J. (2012). The double-edged sword: Does biomechanism increase or decrease judges' sentencing of psychopaths? *Science, 337*, 846–849.

Babor, T., Caetano, R., Casswell, S., Edwards, G., Giesbrecht, N., Graham, K., Grube, J., Hill, L., Holder, H., Homel, R. M. L., Österberg, E., Rehm, J., Room, R., & Rossow, I. (2010a). *Alcohol: No ordinary commodity. Research and public policy*. Oxford: Oxford University Press.

Babor, T., Miller, P., & Edwards, G. (2010b). Vested interests, addiction research and public policy. *Addiction, 105*, 4–5.

Backlar, P. (1996). Genes and behavior: Will genetic information change the way we see ourselves? *Community Mental Health Journal, 32*, 205–209.

BBC News. (2004). 'DNA test' to help smokers quit. *BBC News*. http://news.bbc.co.uk/2/hi/health/4061137.stm. Accessed 25 Jan 2007.

BBC News. (2007). Brain's 'addiction centre' found. *BBC News*. http://news.bbc.co.uk/2/hi/health/6298557.stm. Accessed 2 Aug 2007.

Bell, E., & Racine, E. (2009). Enthusiasm for functional magnetic resonance imaging (FMRI) often overlooks its dependence on task selection and performance. *The American Journal of Bioethics, 9*, 23–25.

Bierut, L. J. (2011). Genetic vulnerability and susceptibility to substance dependence. *Neuron, 69*, 618–627.

Buchman, D. Z., Illes, J., & Reiner, P. B. (2011). The paradox of addiction neuroscience. *Neuroethics, 4*, 65–77.

Burge, T. (2010). A real science of mind. *The New York Times*. http://opinionator.blogs.nytimes.com/2010/12/19/a-real-science-of-mind/?emc=eta1. Accessed 25 Jan 2013.

Campbell, N. D. (2012). Medicalization and biomedicalization: Does the diseasing of addiction fit the frame? In J. Netherland (Ed.), *Critical perspectives on addiction* (Advances in medical sociology, Vol. 14, pp. 3–25). Bingley: Emerald Group.

Cappella, J. N., Lerman, C., Romantan, A., & Baruh, L. (2005). News about genetics and smoking. *Communication Research, 32*, 478.

Carlsten, C., & Burke, W. (2006). Potential for genetics to promote public health: Genetics research on smoking suggests caution about expectations. *Journal of the American Medical Association, 296*, 2480–2482.

Caron, L., Karkazis, K., Raffin, T. A., Swan, G., & Koenig, B. A. (2005). Nicotine addiction through a neurogenomic prism: Ethics, public health, and smoking. *Nicotine & Tobacco Research, 7*, 181–197.

Carter, A., & Hall, W. (2012). *Addiction neuroethics: The promises and perils of neuroscience research on addiction*. London: Cambridge University Press.

Chapman, S., & MacKenzie, R. (2010). The global research neglect of unassisted smoking cessation: Causes and consequences. *PLoS Medicine, 7*, e1000216.

Cheetham, A., Allen, N. B., Whittle, S., Simmons, J. G., Yücel, M., & Lubman, D. I. (2012). Orbitofrontal volumes in early adolescence predict initiation of cannabis use: A 4-year longitudinal and prospective study. *Biological Psychiatry, 71*, 684–692.

Chen, L. S., Baker, T. B., Piper, M. E., Breslau, N., Cannon, D. S., Doheny, K. F., Gogarten, S. M., Johnson, E. O., Saccone, N. L., & Wang, J. C. (2012). Interplay of genetic risk factors (CHRNA5-CHRNA3-CHRNB4) and cessation treatments in smoking cessation success. *The American Journal of Psychiatry, 169*(7), 735–742.

Choudhury, S., Nagel, S. K., & Slaby, J. (2009). Critical neuroscience: Linking neuroscience and society through critical practice. *BioSocieties, 4*, 61–77.

Collins, F. (1999). Medical and societal consequences of the Human Genome Project. *The New England Journal of Medicine, 341*, 28–37.

Condit, C. M., Ofulue, N., & Sheedy, K. M. (1998). Determinism and mass-media portrayals of genetics. *American Journal of Human Genetics, 62*, 979–984.

Conrad, P. (2007). *The medicalization of society: On the transformation of human conditions into treatable disorders*. New York: Johns Hopkins.

Cox, W. M., Hogan, L. M., Kristian, M. R., & Race, J. H. (2002). Alcohol attentional bias as a predictor of alcohol abusers' treatment outcome. *Drug and Alcohol Dependence, 68*, 237–243.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Doidge, N. (2007). *The brain that changes itself: Stories of personal triumph from the frontiers of brain science*. New York: Penguin Books.

Doyle, C. (2004). DNA test can identify 'the smoker's gene'. *The Telegraph*. http://www.telgraph.co.uk/news/main.jhtml?xml=/news/2004/12/02/nfag02.xml&sSheet=/news//12/02/ixhome.html. Accessed 25 Jan 2007.

Draganski, B., Gaser, C., Kempermann, G., Kuhn, H. G., Winkler, J., Büchel, C., & May, A. (2006). Temporal and spatial dynamics of brain structure changes during extensive learning. *Journal of Neuroscience, 26*, 6314–6317.

Duka, T., Crombag, H. S., & Stephens, D. N. (2011). Experimental medicine in drug addiction: Towards behavioral, cognitive and neurobiological biomarkers. *Journal of Psychopharmacology, 25*, 1235–1255.

Edwards, A. C., Svikis, D. S., Pickens, R. W., & Dick, D. M. (2009). Genetic influences on addiction. *Primary Psychiatry, 16*, 40.

Ersche, K., Roiser, J., Lucas, M., Domenici, E., Robbins, T., & Bullmore, E. (2011). Peripheral biomarkers of cognitive response to dopamine receptor agonist treatment. *Psychopharmacology, 214*, 779–789.

Ersche, K. D., Jones, P. S., Williams, G. B., Turton, A. J., Robbins, T. W., & Bullmore, E. T. (2012). Abnormal brain structure implicated in stimulant drug addiction. *Science, 335*, 601–604.

Ersche, K. D., & Robbins, T. W. (2011). An integrated framework for human neuroimaging studies of addiction from a preclinical perspective. In B. Adinoff & E. Stein (Eds.), *Neuroimaging in addiction* (pp. 7–35). Chichester: Wiley.

Evans, J. P., Meslin, E. M., Marteau, T. M., & Caulfield, T. (2011). Deflating the genomic bubble. *Science, 331*, 861–862.

Farah, M. J. (2012). The puzzle of neuroimaging and psychiatric diagnosis: Technology and nosology in an evolving discipline. *American Journal of Bioethics – Neuroscience, 3*(4), 1–11.

Fontenelle, L. F., Oostermeijer, S., Harrison, B. J., Pantelis, C., & Yucel, M. (2011). Obsessive-compulsive disorder, impulse control disorders and drug addiction: Common features and potential treatments. *Drugs, 71*, 827–840.

Garnett, A., Whiteley, L., Piwowar, H., Rasmussen, E., & Illes, J. (2011). Neuroethics and fMRI: Mapping a fledgling relationship. *PLoS One, 6*, e18537.

Gartner, C. E., Barendregt, J. J., & Hall, W. (2009). Multiple genetic tests for susceptibility to smoking do not outperform simple family history. *Addiction, 104*, 118–126.

Geppert, C. M. A., & Roberts, L. W. (2005). Ethical issues in the use of genetic information in the workplace: A review of recent developments. *Current Opinion in Psychiatry, 18*, 518–524.

Goudriaan, A., Oosterlaan, J., De Beurs, E., & Van Den Brink, W. (2008). The role of self-reported impulsivity and reward sensitivity versus neurocognitive measures of disinhibition and decision-making in the prediction of relapse in pathological gamblers. *Psychological Medicine, 38*, 41–50.

Greely, H. (2001). Genotype discrimination: The complex case for some legislative protection. *University of Pennsylvania Law Review, 149*, 1438–1505.

Gu, H., Salmeron, B. J., Ross, T. J., Geng, X., Zhan, W., Stein, E. A., & Yang, Y. (2010). Mesocorticolimbic circuits are impaired in chronic cocaine users as demonstrated by resting-state functional connectivity. *NeuroImage, 53*, 593–601.

Gundle, K. R., Dingel, M. J., & Koenig, B. A. (2010). ‘To prove this is the industry’s best hope’: Big tobacco’s support of research on the genetics of nicotine addiction. *Addiction, 105*, 974–983.

Hall, M. A., & Rich, S. S. (2000). Laws restricting health insurers’ use of genetic information: Impact on genetic discrimination. *American Journal of Human Genetics, 66*, 293–307.

Hall, W. (2005). British drinking: A suitable case for treatment? *British Medical Journal, 331*, 527–528.

Hall, W. (2007). A research agenda for assessing the potential contribution of genomic medicine to tobacco control. *Tobacco Control, 16*, 53–58.

Hall, W., & Chikritzhs, T. (2011). The Australian alcopops tax revisited. *The Lancet, 377*, 1136–1137.

Hall, W., Gartner, C. E., & Carter, A. (2008). The genetics of nicotine addiction liability: Ethical and social policy implications. *Addiction, 103*, 350–359.

Hall, W., Madden, P., & Lynskey, M. (2002). The genetics of tobacco use: Methods, findings and policy implications. *Tobacco Control, 11*, 119–124.

Hall, W. D., Mathews, R., & Morley, K. I. (2010). Being more realistic about the public health impact of genomic medicine. *PLoS Medicine, 7*, e1000347.

Heinrichs, J. H. (2012). The sensitivity of neuroimaging data. *Neuroethics, 1–11*.

Ho, M. K., Goldman, D., Heinz, A., Kaprio, J., Kreek, M. J., Li, M. D., Munafo, M. R., & Tyndale, R. F. (2010). Breaking barriers in the genomics and pharmacogenetics of drug addiction. *Clinical Pharmacology and Therapeutics, 88*, 779–791.

Hutchison, K. E. (2010). Substance use disorders: Realizing the promise of pharmacogenomics and personalized medicine. *Annual Review of Clinical Psychology, 6*, 577–589.

Illes, J., & Racine, E. (2005). Imaging or imagining? A neuroethics challenge informed by genetics. *The American Journal of Bioethics, 5*, 5–18.

Ioannidis, J. P. (2011). Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry*. doi:10.1001/archgenpsychiatry.2011.28.

Ioannidis, J. P. (2012). Genetic prediction of common disease: Will personal genomics ever work? *Archives of Internal Medicine, 172*, 744–746.

Ioannidis, J. P. A. (2013). Biomarker Failures. *Clinical Chemistry, 59*, 202–204.

Ioannidis, J. P., Castaldi, P., & Evangelou, E. (2010). A compendium of genome-wide associations for cancer: Critical synopsis and reappraisal. *Journal of the National Cancer Institute, 102*, 846–858.

Ioannidis, J. P. A. (2009). Limits to forecasting in personalized medicine: An overview. *International Journal of Forecasting, 25*, 773–783.

John, U., Meyer, C., Hapke, U., Rumpf, H. J., & Schumann, A. (2004). Nicotine dependence, quit attempts, and quitting among smokers in a regional population sample from a country with a high prevalence of tobacco smoking. *Preventive Medicine, 38*, 350–358.

Khoury, M. J., Thrasher, J. F., Burke, W., Gettig, E. A., Fridinger, F., & Jackson, R. (2000). Challenges in communicating genetics: A public health approach. *Genetics in Medicine, 2*, 198–202.
Khoury, M. J., Yang, Q. H., Gwinn, M., Little, J. L., & Flanders, W. D. (2004). An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genetics in Medicine, 6*, 38–47.
Klöppel, S., Abdulkadir, A., Jack, C. R., Jr., Koutsouleris, N., Mourão-Miranda, J., & Vemuri, P. (2012). Diagnostic neuroimaging across diseases. *NeuroImage, 61*, 457–463.
Lazar, S. W., Kerr, C. E., Wasserman, R. H., Gray, J. R., Greve, D. N., Treadway, M. T., McGarvey, M., Quinn, B. T., Dusek, J. A., & Benson, H. (2005). Meditation experience is associated with increased cortical thickness. *Neuroreport, 16*, 1893.
Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*, 869–878.
Loth, E., Carvalho, F., & Schumann, G. (2011). The contribution of imaging genetics to the development of predictive markers for addictions. *Trends in Cognitive Sciences, 15*, 436–446.
Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences, 97*, 4398–4403.
Marteau, T. M., French, D. P., Griffin, S. J., Prevost, A., Sutton, S., Watkinson, C., Attwood, S., & Hollands, G. J. (2010). Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours. *Cochrane Database of Systematic Reviews, 10*.
Mathews, R., Hall, W., & Carter, A. (2012). Direct-to-consumer genetic testing for addiction susceptibility: A premature commercialisation of doubtful validity and value. *Addiction, 107*(12), 2069–2074.
McBride, C. M., Bowen, D., Brody, L. C., Condit, C. M., Croyle, R. T., Gwinn, M., Khoury, M. J., Koehly, L. M., Korf, B. R., Marteau, T. M., McLeroy, K., Patrick, K., & Valente, T. W. (2010). Future health applications of genomics: Priorities for communication, behavioral, and social sciences research. *American Journal of Preventive Medicine, 38*, 556–565.
Merikangas, K. R., & Risch, N. (2003). Genomic priorities and public health. *Science, 302*, 599–601.
Midanik, L. (2006). *Biomedicalization of alcohol studies: Methodological shifts and institutional challenges*. New Brunswick: Transaction Publishers.
Miller, P., Carter, A., & De Groot, F. (2012). Investment and vested interests in neuroscience research of addiction: Why research ethics requires more than informed consent. In A. Carter, W. Hall, & J. Illes (Eds.), *Addiction neuroethics: The ethics of addiction research and treatment* (pp. 278–301). New York: Elsevier.
Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 2693–2698.
Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K. A., Mansfield, A., Sinnott-Armstrong, W., & Gazzaniga, M. (2010). Neuroprediction, violence, and the law: Setting the stage. *Neuroethics*. doi:10.1007/s12152-010-9095-z.
Netherland, J. (2011). "We haven't sliced open anyone's brain yet": Neuroscience, embodiment and the governance of addiction. In M. Pickersgill & I. Van Keulen (Eds.), *Sociological reflections on the neurosciences* (Advances in medical sociology, Vol. 13, pp. 153–177). Bingley: Emerald Group.
Parvaz, M. A., Alia-Klein, N., Woicik, P. A., Volkow, N. D., & Goldstein, R. Z. (2011). Neuroimaging for drug addiction and related behaviors. *Reviews in the Neurosciences, 22*, 609–624.
Paulus, M. P., Tapert, S. F., & Schuckit, M. A. (2005). Neural activation patterns of methamphetamine-dependent subjects during decision making predict relapse. *Archives of General Psychiatry, 62*, 761–768.

Pierce, J. P., Gilpin, E. A., Emery, S. L., White, M. M., Rosbrook, B., & Berry, C. C. (1998). Has the California tobacco control program reduced smoking? *Journal of the American Medical Association, 280*, 893.
Pievani, M., de Haan, W., Wu, T., Seeley, W. W., & Frisoni, G. B. (2011). Functional network disruption in the degenerative dementias. *Lancet Neurology, 10*, 829–843.
Press, N. (2006). Social construction and medicalization: Behavioral genetics in context. In E. Parens, A. R. Chapman, & N. Press (Eds.), *Wrestling with behavioral genetics: Science, ethics and public conversation* (pp. 131–149). Baltimore: Johns Hopkins University Press.
Quart, A. (2012). Neuroscience: Under attack. *The New York Times*. http://www.nytimes.com/2012/11/25/opinion/sunday/neuroscience-under-attack.html?_r=0. Accessed 25 Jan 2013.
Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews. Neuroscience, 6*, 159–164.
Racine, E., Bar-Ilan, O., & Illes, J. (2006). Brain imaging: A decade of coverage in the print media. *Science Communication, 28*, 122–142.
Reske, M., & Paulus, M. P. (2011). The diagnostic and therapeutic potential of neuroimaging in addiction medicine. *Neuroimaging in Addiction, 319–343*.
Rose, N. (2006). *The politics of life itself: Biomedicine, power, and subjectivity in the twenty-first century*. Princeton: Princeton University Press.
Rose, N. (2010). 'Screen and intervene': Governing risky brains. *History of the Human Sciences, 23*, 79–105.
Roskies, A. L. (2008). Neuroimaging and inferential distance. *Neuroethics, 1*, 19–30.
Rothenberg, K., Fuller, B., Rothstein, M., Duster, T., Kahn, M. J. E., Cunningham, R., Fine, B., Hudson, K., King, M. C., Murphy, P., Swergold, G., & Collins, F. (1997). Genetic information and the workplace: Legislative approaches and policy challenges. *Science, 275*, 1755–1757.
Rothstein, M. A., & Anderlik, M. R. (2001). What is genetic discrimination, and when and how can it be prevented? *Genetics in Medicine, 3*, 354–358.
Schleim, S., & Roiser, J. P. (2009). FMRI in translation: The challenges facing real-world applications. *Frontiers in Human Neuroscience, 3*, 63.
Schüll, N. D. (2006). Machines, medication, modulation: Circuits of dependency and self-care in Las Vegas. *Culture, Medicine and Psychiatry, 30*, 223–247.
Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., Conrod, P. J., Dalley, J. W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J. L., Paus, T., Poline, J. B., Robbins, T. W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D. N., Strohle, A., & Struve, M. (2010). The IMAGEN study: Reinforcement-related behaviour in normal brain function and psychopathology. *Molecular Psychiatry, 15*, 1128–1139.
Schutz, C. (2008). Using neuroimaging to predict relapse to smoking: Role of possible moderators and mediators. *International Journal of Methods in Psychiatric Research, 17*, S78–S82.
Schweitzer, N. J., Saks, M. J., Murphy, E. R., Roskies, A. L., Sinnott-Armstrong, W., & Gaudet, L. M. (2011). Neuroimages as evidence in a mens rea defense: No impact. *Psychology, Public Policy, and Law, 17*, 357.
Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., & Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron, 62*, 42.
Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature, 460*, 202–207.
Sinha, R. (2011). New findings on biological factors predicting addiction relapse vulnerability. *Current Psychiatry Reports, 13*, 398–405.
Sinnott-Armstrong, W., Roskies, A., Brown, T., & Murphy, E. (2008). Brain images as legal evidence. *Episteme, 5*, 359–373.
Sutherland, M. T., McHugh, M. J., Pariyadath, V., & Stein, E. A. (2012). Resting state functional connectivity in addiction: Lessons learned and a road ahead. *NeuroImage, 62*, 2281–2295.
Swendsen, J., & Le Moal, M. (2011). Individual vulnerability to addiction. *Annals of the New York Academy of Sciences, 1216*, 73–85.

van Praag, H. (2009). Exercise and the brain: Something to chew on. *Trends in Neurosciences, 32*, 283–290.
Verweij, M. (1999). Medicalization as a moral problem for preventive medicine. *Bioethics, 13*, 89–113.
Volkow, N. D., & Li, T.-K. (2005). Drugs and alcohol: Treating and preventing abuse, addiction and their medical consequences. *Pharmacology & Therapeutics, 108*, 3–17.
Vos, T., Carter, R., Barendregt, J., Mihalopoulos, C., Veerman, L., Magnus, A., Cobiac, L., Bertram, M., & Wallace, A. (2010). *Assessing cost-effectiveness in prevention: ACE–prevention September 2010 final report*. Brisbane: The University of Queensland.
Vrecko, S. (2008). Capital ventures into biology: Biosocial dynamics in the industry and science of gambling. *Economy and Society, 37*, 50–67.
Whelan, R., Conrod, P. J., Poline, J.-B., Lourdusamy, A., Banaschewski, T., Barker, G. J., Bellgrove, M. A., Buchel, C., Byrne, M., Cummins, T. D. R., Fauth-Buhler, M., Flor, H., Gallinat, J., Heinz, A., Ittermann, B., Mann, K., Martinot, J.-L., Lalor, E. C., Lathrop, M., Loth, E., Nees, F., Paus, T., Rietschel, M., Smolka, M. N., Spanagel, R., Stephens, D. N., Struve, M., Thyreau, B., Vollstaedt-Klein, S., Robbins, T. W., Schumann, G., & Garavan, H. (2012). Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience, 15*(6), 920–925.
White, V., Hill, D., Siahpush, M., & Bobevski, I. (2003). How has the prevalence of cigarette smoking changed among Australian adults? Trends in smoking prevalence between 1980 and 2001. *Tobacco Control, 12*, ii67.
Willett, W. C. (2002). Balancing life-style and genomics research for disease prevention. *Science, 296*, 695–698.
Wright, A. J., Weinman, J., & Marteau, T. M. (2003). The impact of learning of a genetic predisposition to nicotine dependence: An analogue study. *Tobacco Control, 12*, 227–230.
Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: Neuroimaging changes in brain structure during learning. *Nature Neuroscience, 15*, 528–536.

Ethical Issues in the Treatment of Addiction 67

Benjamin Capps, Adrian Carter, and Yvette van der Eijk

Contents

Abstract

Addiction or dependence to a psychoactive substance – or drug – is a biopsychosocial disorder. Many aspects of addiction can be clinically treated, yet this raises some important ethical issues. The health and social problems resulting from addiction are generally influenced by sociopolitical agendas and

B. Capps (✉) • Y. van der Eijk
Centre for Biomedical Ethics, National University of Singapore Yong Loo Lin School of Medicine, Singapore
e-mail: medbjc@nus.edu.sg; y.vandereijk@nus.edu.sg

A. Carter
The University of Queensland, UQ Centre for Clinical Research, Royal Brisbane and Women's Hospital, Herston, QLD, Australia
e-mail: adrian.carter@uq.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_70,

prevailing views about addiction, and therefore, we make some brief observations about how these affect the act, manner, or method of treating someone who is addicted.

In this chapter, we focus on some recent developments in the treatment of addiction: pharmacological approaches, and invasive brain therapies. In addition to giving an ethical overview of these, we look more generally at the justification for coerced treatment.

Introduction

This chapter considers the ethical aspects of certain treatments of addiction. Our focus is on psychoactive substances – more commonly known as drugs – that can produce an addictive pattern of use.

By addiction, we refer to a biopsychosocial disorder that is indicated by seemingly uncontrollable cravings for a particular substance (Engel 1980). The term “biopsychosocial” refers to the observation that substance dependence can result from (a combination of) various factors such as neurological adaptations, comorbid psychological disorders, or social circumstances. In this chapter, we will use “dependence” and “addiction” interchangeably to signify typical behavior associated with an addicted state. The American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association 1994) terms addiction as “substance dependence” and diagnoses it by three or more of the following criteria within a 12-month period: (1) tolerance; (2) withdrawal; (3) consuming more than intended; (4) persistent desire or unsuccessful effort to cut down; (5) spending a great deal of time with drug-related activity; (6) giving up important social, occupational, or recreational activities; and (7) continued consumption despite physical or psychological harm. While revisions of this definition are expected in DSM-V, due for release in 2013 (American Psychiatric Association 2012), defining such an intricate disorder has proved to be problematic. And yet such attempts underlie the ethical, political, and legal aspects of addicted states; the way in which it is perceived has important ethical implications for dependent individuals, those closely associated with them, and society as a whole.

The health and social problems resulting from addiction are generally influenced by sociopolitical agendas and the prevailing partisan and scientific understandings of addiction; such considerations might be scientifically informed, inclusive of current knowledge, or narrowly defined by prejudice or bias. How one thinks about addiction, therefore, will define the terms of the act, manner, or method of “treating” someone who is addicted. It is our intention to remain entirely neutral and non-prejudiced with respect to the terminologies “addict,” “person with addiction,” “drug user,” and similar definitions. Since our focus is on treatments, often we will also refer to the “patient.” Similarly, we refer to “drug use” to indicate the ingestion of a substance known to be associated with “addicted-like states” (*qua* “addictive drugs”). We avoid the

term “misuse” because this refers to the illegality (that may depend on jurisdiction) and essentially irresponsible ingestion of, and actions associated with, addictive drugs.

At opposing apexes are two prevailing models of dependence: the “moral choice” and the “medical” models. The moral choice model represents the conventional view in which individuals freely choose to engage in drug use. On this view, the ostensible state of neurological dependence is primarily used as an excuse for substance abuse. "Addictive behavior, therefore, is principally a deficit of the will (because of the choice to use – or not use drugs), and this justifies a punitive approach to hold such individuals responsible for misuse and any activities associated with it. Addicts may be subject to punishment and forced abstention, rather than voluntary therapy – a mainstay of traditional medical ethics. The medical model, on the other hand, while not excusing choice in all circumstances of drug use, does challenge assumptions about an addicted state: based on recent advances in neurobiological research, addiction can be viewed as a medical or more specifically, psychiatric condition that can be treated by evidence-based medical interventions. The medical model of addiction has led to the more specific chronic and relapsing brain-disease model of addiction (Koob and Le Moal 2006).

The aim of this chapter is not to take a stand on a specific ideology of addiction or approach to treatment but to discuss the ethical implications of implementing certain views in regard to interventions in addiction. Our approach has two prongs: one, to identify the relevant empirical literature that is used to ground opinion and policies in addiction treatment and two, to provide an introduction to some of the theoretical ideas that are supported by these observations. Admittedly, there will be gaps in our analysis, but our hope is that the reader will become aware of the current state of debate in addiction treatment. This involves understanding a complex interplay between the recognition of symptoms of dependence – whether biological and/or behavioral; attitudes to choices that lead to use, abuse, and misuse; experiences of harm caused by dependence; and partisan views of addiction, often swayed by political agendas, economics, and morals. Thus, the availability and means by which treatment is either offered or compelled is predicated by the view adopted. In this respect, we offer some suppositions in respect to the ethical issues in the following areas of addiction therapy: (1) the pharmacological treatment of drug addiction; (2) invasive therapies to reduce or treat addiction, such as deep brain stimulation and neurosurgery; and (3) more generally, the justification for coerced treatment.

The Pharmacological Treatment of Drug Addiction

To understand treatment approaches, one has to consider a number of factors: the availability and acceptance of evidence-based treatments, the wider causes of dependence (the clinical, or otherwise, diagnosis, including the social and biochemical triggers), and what constitutes successful – and the goals of – treatment. Much of the tension in treatment programs arises from

a difference between two broad approaches to drug dependence: (1) approaches that aim at achieving abstinence in the short to medium term and (2) harm minimization approaches which see abstinence as a long-term goal and aims in the short to medium term to encourage less harmful forms of drug use in the interests of reducing harm to users and the community. While these two approaches are sometimes characterized as mutually exclusive by some of their proponents, they are more accurately seen as lying on a continuum of treatment options. Many of the ethical issues in treatment of, for example, opioid dependence, arise from this difference in approach. How individuals align themselves within this debate is often driven by their understanding of the nature of dependence and by assumptions about the degree of autonomy and responsibility that individuals have over their conduct. In this first section, we comment on a wide spectrum of broadly pharmacological treatments.

Drug Substitution, Maintenance, and Relapse Treatments

Long-term pharmacological approaches to treating addiction can be understood as falling in to one of the two categories: (1) maintenance or substitution treatments that aim to reduce the harm caused by addiction by administering a safer drug that mimics the effect of the abused drug (referred to as an agonist) and (2) relapse prevention programs that aim to achieve and maintain abstinence following detoxification. Although this framework for understanding addiction treatment may not necessarily reflect the more varied, dynamic approaches in the clinical setting (see Carter and Hall 2012), treatment is premised on the grounds that it is possible to not only address the health of the patient but also control access to illegal substances and therefore minimize the potentially unlawful activities and the consequences associated with criminal (mis/ab)use.

In drug substitution or maintenance, a patient is prescribed an agonist that mimics the effects of the current drug of abuse but with a safer pharmacokinetic profile. Maintenance therapies are designed to stabilize the lives of individuals and reduce the harmful activities associated with their current drug use (e.g., overdose or comorbid disease), usually with the long-term aim of weaning individuals off addictive drugs altogether. One example is methadone, an opioid agonist that is legally taken (when it is prescribed) within a treatment program and has been effectively used as a substitution for heroin and morphine for almost five decades – principally by reducing cravings caused by drug abstention. Studies have also shown that treatment for heroin addiction, such as methadone maintenance therapy, results in decreased criminal and violent behavior (Ward et al. 1998). Nicotine replacement therapy (NRT) is an example of a maintenance therapy which aims to reduce the health-related harms of tobacco. Substitution drugs for psychostimulants and cannabis are currently being sought.

Relapse prevention often involves the prescription of a drug that blocks or attenuates the desired effects of the drug of abuse (e.g., Antabuse or disulfiram

for alcohol dependence or naltrexone for opioid dependence). The aim of relapse prevention treatments is to prevent a single use of a drug – a "slip" – leading to a relapse to chronic drug use.

Both approaches have been criticized on clinical grounds. For example, naltrexone is an opioid receptor antagonist used primarily in the treatment of alcohol or opioid dependence. It has been suggested that its long-term use may produce dysphoria and depressive symptoms (Dean et al. 2006) and attenuate the rewarding or hedonic effects of everyday "authentic" activities such as eating (Yeomans and Gray 2002), sex (Murphy et al. 1990), and physical exercise (Daniel et al. 1992). Agonists, such as methadone, are themselves addictive and can cause harms similar to those they are used to replace (e.g., overdose).

However, there is also often a moral refutation of such treatments, and which is preceded by a long history of moralizing about drug use, and policy and political attempts to curtail drug access because of social and other ostensible harms. This has led to a complex debate about the categorization of drugs in terms of opposing theories of liberty or freedom and moral conservatism; the latter of which can be used to label drug misuse as immoral. While unpacking these moral theories is beyond this chapter, one pertinent concern we can look briefly at is that these notional *treatment* strategies affect behavior in similar ways to illegally abused or misused drugs. (An issue here, but which we will have to leave unexplored, is whether a drug is categorized as illicit because of the harms it causes, its social acceptance, or some other artifact that has more dubious justification). So, even if in clinical terms these approaches can reduce the expected harm to the individual's health and intercedes in their drug-seeking activities, their application also contribute to the agent's loss of moral responsibility, and therefore, providing addicted individuals with addictive drugs (or with drugs with similar effects) is not treatment at all.

If one considers drug use to be morally contemptible because of the contemptuous effects it has on the individual, then it is unlikely that inducing similar effects for medical reasons is going to be ethically acceptable; the "excusive" reasoning used, it might be argued, becomes regressively circular. The solution is to denounce the causal factor of the dissipated will – drug misuse. So, in this respect, if goals and our realization of them are indications of our "sense of self" – and authentic behavior is part of our essential wellbeing – then offering alternative substances that merely perpetuate this deficit is just as ethically contentious. As some bioethicists have noted, all drugs that interfere with or undermine this sort of authentic action should be rebuked (Kass 2003). In this respect, some have opined that although less harm is expected by treating addiction by means of substitution or relapse prevention, they are still responsible for compromising the agent's will and thus do not stabilize this deficit. Moreover, "treating" addiction with another, albeit less harmful drug but which could also result in addiction, is speculatively not medically indicated (supposing that medicine's purpose is to make a patient well, then supplying potentially addictive drugs should be precluded).

In contrast, while the use of drugs by healthy adults to alter their sense of self may be an important neuro-ethical concern (Levy 2007), the situation is arguably very different in the case of addicted individuals who already have a distorted concept of

self – these individuals might be considered as *ill*. As such, addiction may significantly affect their will and resulting opportunities, and constrain the kind of life that is open to them: the illness produces significant emotional, physical, and psychological harms that limit their choices. A drug that is able to ameliorate these personal conditions could more correctly be seen as enabling, rather than impairing, the effected will.

In this brief analysis, one can clearly see how the construal of addiction – either as an undesired neurobiological manifestation or a personal (but disapprobational) choice – can influence treatment options and opportunities. On the one hand, if a person is considered to be taking drugs out of choice, then they do not have an impaired capacity to make decisions – therefore no treatment of this kind is warranted and stopping access to the drug altogether is a justified intervention. On the other hand, if they are medically unwell, then any means of recovery would be justified.

More generally, however, is a concern that a categorical rejection of the therapeutic use of any one of these approaches (for whatever reason) may lead to the sole use of another – sometimes the alternatives are less clinically effective, or, in the case of invasive operations on the brain (as we shall discuss below), less welcome. For example, the antagonist naltrexone may be used to treat opioid dependence at the expense of more effective treatments such as methadone and buprenorphine (agonists) – both which are prohibited in Russia (Krupitsky et al. 2010) and only offered in a limited manner in many other jurisdictions. Previous experience with substitute prescribing for nicotine dependence reveals a perverse regulatory standard that insists upon much tighter regulations for less harmful nicotine products, such as smokeless tobacco (e.g., snus), than are imposed on smoked tobacco products (Gartner et al. 2007). The reason for withholding some treatments is again an issue that is grounded in the historical and social context of drug use (Courtwright 2001). In a related example, a safe substitute for alcohol has been suggested in the form of a GABA partial-agonist that enjoys the social properties of alcohol without its biological toxicity (Nutt 2006). The idea, however, elicits the kinds of opinions that already call for regulatory obstacles to these “fringe” solutions and “engineered alternatives”; they only increase the range of recreational drugs available and the science does not address the “authenticity” concerns of drug misuse (or indeed the perceived social “harms” of recreational drug use). Such attitudes provide a major disincentive to pharmaceutical interest and political investment in harm reduction approaches. Although it is possible that safer recreational drugs will emerge as a by-product of basic pharmaceutical research, it is unlikely that there would be invested appropriate funds or that an effective political will would emerge for similar solutions to recreational drugs.

Long-Acting Treatment of Addiction: Immunological Blockade and Drug Implants

Researchers are seeking technological innovations to overcome the failure of addicted individuals to *comply* with pharmacological treatment by voluntarily

administering long-acting treatments. Although, as we shall discuss below, their coercive use may well signify an ambiguity with respect to choice and drug use – vaccines and other such substances may remove "choice" altogether. In respect to clinical use, long-acting treatments could be viewed in one of the two ways: (1) as a form of treatment that imposes a state of compliance (see our discussion on coercion in section "Coerced Addiction Treatment," below) or (2) a treatment that enables addicted individuals to adhere to their higher-order (and often presumed more authentic) desires to remain abstinent – a kind of "Ulyssian" contract to achieve abstinence despite "inauthentic" cravings that are part of their addiction. These treatments include long-acting interventions such as drug vaccines and drug implants, which, unlike oral treatments, cannot be easily circumvented or discontinued to allow individuals to regain the prior desired effects of drug use. Unlike traditional pharmacological treatments that require daily dosing, long-lasting treatments only need to be taken every 1–6 months (Harwood and Myers 2004).

"Vaccines" are a novel approach to tackling addiction. As prophylactics, they may prevent addiction by neutralizing a drug's rewarding effect in, for example, adolescents in their initial stages of drug use and experimentation. However, this application has failed to gain much traction for a number of reasons (e.g., Ashcroft et al. 2007). Vaccination could also be used as an addiction therapy to alleviate the effects of drug withdrawal and prevent possible relapse in persistent drug users. Although vaccines have been developed against methamphetamine, opioids, nicotine, cocaine, and phencyclidine (Orson et al. 2008), the majority of research has focussed on nicotine and cocaine.

The nicotine vaccine induces nicotine-specific antibodies that bind to the target drug preventing it from crossing the blood–brain barrier and acting on receptors in the brain (Nutt and Lingford-Hughes 2004). According to animal studies, drug vaccines reduce the amount of drug that reaches the brain and thereby can reduce the self-administration of the target drug. They are also highly specific and, because they do not cross the blood–brain barrier, produce no adverse effects on the central nervous system (Kosten and Owens 2005). This makes the nicotine vaccine an attractive alternative to drugs such as bupropion and varenicline, which have been associated with depression and suicidal ideation. Vaccination against nicotine could help abstinent smokers during the first few months after quitting when they are most vulnerable to relapse (Vocci and Chiang 2001).

So far, a number of phase II and III clinical trials for cocaine and nicotine vaccines have been reported (e.g., Maastricht University Medical Centre 2009; Martell et al. 2009). Only one-third of participants developed sufficient antibodies against the drug to prevent a relapse. Those individuals that developed the highest level of antibodies derived the most benefit from the intervention. Increasing the dosage is a possibility, but this could also increase the chance of potential side effects. A phase III clinical trial of the NicVax nicotine vaccine found it to be ineffective. New formulations of vaccines are currently in development that may provide greater immunoprotection against relapse (Moreno et al. 2010), but for now their practical utility remains elusive.

As an alternative to prophylactic vaccines, slow-release formulations are being developed. They typically come in the form of depot injections – oil suspensions that are injected into the muscle and drug implants – larger polymer-based implants inserted under the skin that slowly release the medication over weeks and months.

Researchers have focussed on the development of long-acting formulations of antagonist drugs that block or attenuate the use of addictive drugs. Several slow-release preparations of the antagonist naltrexone, used to treat alcohol and opioid dependence, have been developed (see Krupitsky and Blokhina 2010). The USA Food and Drug Administration approved the use of a naltrexone implant (Vivitrol) for the treatment of alcohol in 2006 and for opioid dependence in 2010. Slow-release agonists and partial agonists are also in development, although there is generally less eagerness to offer slow-release drugs closely related to those which are commonly abused. The exception is buprenorphine, a partial opioid agonist, of which several slow-release preparations were trialed for the treatment of opioid dependence (e.g., Ling et al. 2010). In this respect, a recent position statement by the Australian National Council on Drugs recommends that "given the very limited Australian data and evidence on the efficacy and safety of sustained release naltrexone preparations, their authorised use. . . is ethically problematic as it puts patients at risk of unknown harms, for an unknown benefit" (ANCD 2012, s. IX).

Although it remains to be seen whether blocking the rewarding effects of a particular drug is a clinically and socially acceptable way to reduce its use, these technologies also raise a number of ethical concerns in addition to questions over their efficacy. First, vaccines and sustained-release drugs may prove counterproductive if an individual attempts to overcome their blocking effects by increasing their drug dose. For example, one study for a cocaine vaccine found ten times more cocaine in the systems of their participants than researchers had encountered prior to the inoculation (Martell et al. 2009). Drugs such as nicotine and cocaine cause significant physical damage outside of their impact on the central nervous system (e.g., cardiovascular disease), and any increase in use has the potential to be harmful. Moreover, those who ambivalently agree to vaccination may also switch to using other possibly more dangerous drugs, more harmful routes of administration (e.g., intravenous injection), or much higher doses than usual, thus risking overdoses (Murray 2004); this was already seen in the cocaine vaccine trials just mentioned. It is also possible that readily available and effective vaccines could make experimentation with drugs seem less risky – that the possibility of addiction is not a concern for someone "vaccinated" – and therefore unwittingly increase drug use.

Second, the use of a vaccines or implants may also block the action of agonists or partial agonists (e.g., methadone and buprenorphine for opioid dependence). This would prevent the use of substitution therapies while vaccination remains effective and could raise concerns especially if drug substitution therapy turns out to be more effective. Vaccines and implants may also block the action of

medications used in the treatment of other medical conditions, such as opioid analgesics for pain relief (Ashcroft and Franey 2004).

Third, vaccines, implants, and other medical interventions could "medicalize" the socioeconomic problems that often drive drug use and addiction. Drug-based approaches tend to avoid the psychological characteristics underlying the addictive condition, such as compromised decision making and loss of opportunities, as well as comorbid conditions or the social conditions that may have led one into substance abuse in the first place. Vaccinating a school cohort because of fears of drug use does not deal with the circumstances of impecuniousness school environments or peer pressure, for example. A potentially more effective alternative for people in circumstances of social deprivation – a known factor in problematic drug use – is arguably a psychosocial approach, principally involving social interventions such community investment and the creation of opportunities.

Lastly, there are people that genuinely want to abstain from drug use but cannot control their cravings without medical help; for them, medical therapies, whether in the form of agonists, antagonists, or vaccines, should be made available on the basis of what is most effective. The challenge will be to identify those most likely to respond to different forms of treatment. Thus, it is important to neither medicalize nor trivialize addiction as a medically treatable disorder but to distinguish the individuals that would benefit from medical therapy from those who would not. This distinction will depend on neurobiological, as well as psychological and social factors. Thus, while novel therapies such as vaccines may find a place in the management of addiction, they are unlikely to benefit everyone. Addiction is a complex biopsychosocial disorder that deserves more than a "one-size-fits-all" approach. It is this observation that we carry through in the following sections.

Invasive Treatments of Addiction: Deep Brain Stimulation and Neurosurgery

The use of physical interventions to alter the suspected neurobiological centers of psychiatric disorders has an ignominious history. However, neuroscientific enthusiasm has seen a renewed interest in the use of these invasive techniques to treat intractable psychiatric conditions, including addiction (Carter and Hall 2012). The use of these treatments presupposes an illness rather than a mere deficit of will, although in more archaic times, one might have used such interventions to correct that deficit.

Neurosurgery

A novel and currently rarely used intervention for addiction is stereotactic ablative neurosurgery of brain structures thought to be involved in the condition.

Neurosurgery is an invasive and permanent intervention and is generally considered a treatment of last resort in cases of severe, treatment-resistant psychiatric disorders.

Russian and Chinese surgeons have used neurosurgical procedures to treat heroin addiction: 305 patients were reportedly operated on in Russia and over 500 in China. In China, stereotactic surgery was used to destroy the nucleus accumbens (NAc) (Gao et al. 2003), the brain region where rewarding effects of opioids and other drugs are thought to be mediated. Russian neurosurgeons lesioned the cingulate gyrus (CG) – a brain region that has previously been removed to treat obsessional disorders. International criticism prompted both countries to abandon the controversial treatment in 2003; but a recent report suggests that clinicians in China have commenced a clinical trial of neurosurgical treatment of opioid addiction and have trialed the same procedure in alcohol dependent individuals (Wu et al. 2010).

These reports raise a number of ethical concerns. Firstly, since there are effective nonsurgical treatments available for heroin addiction, such as methadone or buprenorphine, there ought to be important medical reasons to supersede established treatments with experimental neurosurgery. According to the clinical concept of *equipoise* – a principle which requires a genuine uncertainty in the medical community of the therapeutic merits of a treatment regimen – it is unethical to trial an invasive and experimental technology where effective and easily provided treatments are available. However, the reasons for not providing what are thought elsewhere to be effective treatments appears to be a sociopolitical and regulatory decision: methadone maintenance therapy was not made available in China at that time, and its use is prohibited by law in Russia.

Secondly, there are major concerns about the safety and efficacy of these neurosurgical procedures. Stereotactic neurosurgery is an invasive and irreversible procedure that involves drilling holes in the patient's skull and inserting electrodes into the brain to destroy the target region. Advocates of these procedures argue that they are less invasive and destructive than older forms of psychosurgery and report low rates of complications. However, these results come from uncontrolled studies that did not properly evaluate the cognitive and behavioral effects of destroying important neurological structures such as the NAc and aCG (Medvedev et al. 2003). Unknown extents of damage in turn leads to major concerns about the effects of producing irreversible lesions in neural centers that are implicated in the control of appetite, sexual behavior, and the formation of social bonds. At particular risk are permanent changes to the patient's responsiveness to reward their motivation and mood states, leading to clinical etiologies.

Finally, there are also doubts about whether patients have given free and informed consent because of the circumstances in which neurosurgery has been used. Chinese policies towards opioid dependence are largely punitive and retributive, with imprisonment and compulsory detoxification as first-line responses. In such circumstances, it is difficult to obtain free and informed consent without being coercive, especially given the drastic nature of neurosurgery and the lack of options for nonsurgical therapies (we talk more about this, below).

Deep Brain Stimulation

Deep brain stimulation (DBS) is a form of neurosurgery that has been proposed as a treatment of addiction (Krack et al. 2010). It involves inserting electrodes into specific brain regions to modulate neural activity via a battery-controlled external stimulator in the patient's chest and has been used to treat resistant cases of Parkinson's disease (PD) for almost three decades. It is also being trialed in intractable psychiatric conditions including Tourette's syndrome, obsessive-compulsive disorder (OCD), and depression (Krack et al. 2010).

DBS is often described as a reversible alternative to neurosurgery: it is nonetheless an invasive intervention that carries significant risks. Estimates of surgical complications vary markedly, largely due to differences in the competence of the teams performing the operation and variations in the procedure. A recent meta-analysis estimated that approximately 11 % of patients have adverse events from surgery (Krack et al. 2010). Estimates of major adverse surgical outcomes, such as intracerebral hemorrhages and death range from 0 % to 10 %, although in most centers, the prevalence of intracerebral hemorrhage is probably less than 2 %. The insertion of stimulating electrodes can also cause infections and produce cognitive, behavioral, or emotional disturbances and irreversible psychosocial harm. Given the risks associated with any neurosurgery, there needs to be a careful assessment of the risks and benefits of the procedure before DBS should be used to treat addiction.

So far, there have been several reports of the use of DBS in humans to treat addiction to nicotine, alcohol, and heroin (for a review, see Carter and Hall 2011). In one study, a woman was unsuccessfully treated for agoraphobia by bilateral DBS of the NAc but incidentally ameliorating her comorbid alcohol dependence. The same group reported smoking cessation in three of ten patients who underwent DBS of the NAc for Tourette's syndrome, OCD, or anxiety, but over two-thirds derived little benefit. A 47-year-old woman treated with DBS of the NAc for treatment-refractory OCD quit smoking and lost weight post-surgery. However, these changes emerged 10 months after her OCD symptoms disappeared, suggesting that this may have been an indirect effect of successful treatment of her OCD. Craving for alcohol and alcohol consumption were greatly reduced in three long-term, treatment refractory alcohol-dependent individuals who underwent DBS of the NAc; two were abstinent after 1 year and a third had markedly reduced their drinking. These were however small case studies with short-term follow-up and no comparison group (Carter and Hall 2011). The history of neurosurgical treatment in psychiatry cautions against uncritically accepting "positive results" from uncontrolled and often selectively reported clinical case series (Schlaepfer and Fins 2010).

Put together, the evidence to support trials of DBS in addiction is poor. The use of DBS in debilitating conditions such as PD is justified by the severity of the condition and the irreversible deterioration in motor function that characterizes the disease. Addiction, while often a serious condition, does not necessarily follow an inexorable path to severe disability and death; it is generally more amenable to pharmacological and psychotherapeutic treatment, and therefore drastic remedies

are less justifiable. Many cases of addiction treatment failure are due to inadequate access to well run and optimally provided forms of existing treatments (Carter and Hall 2012); a situation that would be exacerbated by an increased use of DBS to treat drug addiction. This suggests that the very uncertain benefits of DBS in alleviating the symptoms of addiction do not outweigh the known harms associated with the procedure or the harm of not providing DBS and relying upon currently available treatments provided to the highest standard.

There are other ethical and social issues that would need to be considered in evaluating the use of DBS for addiction. DBS is an extremely expensive procedure, costing about US $50,000 for the operation with US $10,000 ongoing maintenance costs every few years. This would utilize scarce health resources to treat a very small number of addicted patients with the income to pay for it, while failing to treat the majority. This raises a question of distributive justice, especially given the fact that many addicts are of low-income social groups. It may also negatively affect the provision of DBS to patients with illnesses that may benefit more from DBS, such as PD. The opportunity costs of providing DBS, even if it proved to be safe and effective, make such trials a low priority for public funding (Carter and Hall 2011).

One final, but important, consideration is that of patient expectations. Addicted individuals, as well as their loved ones, are often desperate to find a permanent cure and may have unreasonably high expectations based on uncritical media reports, as has happened with previous psychosurgical developments (Diefenbach et al. 1999). Therefore, the management of patient expectations about the limited and uncertain benefit of the treatment is an important challenge. It is important that the patients or their proxies understand that DBS, as well as neurosurgery, are invasive and risky procedures that do not necessarily offer a cure for addiction. Moreover, they cannot eliminate comorbid psychiatric issues or the social circumstances that led to drug use in the first place, and so patients may still require ongoing psychosocial support or social readjustment.

Coerced Addiction Treatment

Punitive Justifications for Coerced Addiction Treatments

One of the greatest challenges in addiction treatment is persuading individuals to enter and remain in treatment. Such individuals engage in risky behaviors that may be harmful to themselves and to others and in some cases also criminal. In many circumstances, these behaviors would appear to be directly related to their addiction. The person may consider themselves as unwell, but they are unable to seek out aid or sustain treatments. Or their condition may deteriorate despite their own efforts. Society may also want to reduce or limit the opportunities for these kinds of behavior and impose conditions and seek redress as a result of harmful acts as a result of drug use. The justification for coercion – to make somebody do something against his or her will by using force or threats – therefore features prominently in these treatment debates. Given the disputed nature of addiction and

the complex responses to it – including medical, social, legal, and political dimensions – we can only scratch the surface of this debate here.

Many countries have used some form of coercion to deliver addiction treatment. All of the treatments we have mentioned – drug substitution and relapse prevention, immunological blockade, and invasive brain manipulation – may be used under conditions of coercion. Coercion, here, falls under the conditions of forced choice in three interrelated aspects: compulsoriness (the person has limited option to refuse), desirableness (the choices offered are ones they would not normally make), and threat (failure to make or maintain a choice could be rebuked). The circumstances of coerced treatment become even more multifaceted when one factors in the possibility of compromised agency, such as situations when drug-using individuals are "out of it."

For our purposes, we can exclude the latter conditions because the person may often be treated in their "best interests" (or possibly supervised or restrained until able to consent) – at that time, they simply cannot signal their refusal to undergo treatment and therefore cannot be coerced in the same way as someone who is competent. We will also exclude the use of direct violence in situations of coercion because such obvious threats and actions (such that might be exercised on non-consenting prisoners) present no opportunity for consent. Such conditions are likely to be used within circumstances of incarceration and for tortuous or punishing affect and are only to be (controversially) justified if provocation of the legal system mandates such a response (and we have no intention here of raising further aspects of the jurisprudential conditions qualifying the punishment of drug users). We can also exclude deception, because such action is justified only by the toleration or acceptance of policies which permit the use of falsehood and misrepresentation to elicit a (fraudulent or misinformed) response or compliance. In coercive circumstances, then, the recipient gives his or her consent but is compelled in a way parallel to forced choice: actions under these conditions are not fully voluntary.

Coercion may be justified under either the moral or medical models of addiction. If coerced therapy leads to better outcomes for the addict *or* society, it may be justified on the grounds of paternalism (i.e., it elicits individual benefits to the patient even though those benefits might not be welcome) or public protection (i.e., purely societal benefits). This may be pivoted between the threat the addict imposes to himself or herself and society. However, if retribution is a priority, the moral choice model might be used to force unwanted interventions upon a convicted criminal on punitive grounds (in the sense that such interventions are not offered voluntarily because the convict might indulge "medicalized" excuses).

On the one hand, in failing to comply with the law (and in making that choice), the agent sacrifices their control over future circumstances so that an intervention can be compelled as part of rehabilitation or punishment. A person might also be offered a less punitive sanction if they agree to "treatment," but treatment is likely to be part of a vexed intention to deal with other harms such as HIV transmission. On the other hand, the medical model might use coercion as a veiled "best interest" justification so that nonconsensual treatment may be clinically justified or to offer alternatives as "the public interests," as in the case of drug courts where the

justification may be the benefits of rehabilitation rather than retribution. In the punitive setting, treatment may be offered as a *restrained* choice, i.e., an alternative to other sanctions. Such an approach has become commonplace in the USA with the emergence of Diversionary Drug Courts. Similar "progressive" programs were enthusiastically adopted by countries such as Australia, Canada, and the UK in the absence of strong empirical evidence demonstrating their efficacy (Carter and Hall 2012).

Not all jurisdictions operate the same; but basically, if the defendant qualifies for trial in one of these courts (depending on jurisdiction, type of offence, and the involvement of drug in committing the offence), and after an admitting plea, they may be given a suspended or some lesser sentence conditional on the successful completion of the treatment program. Courts can also impose a series of sanctions and can award privileges during that time. If the program is not completed successfully, the participant returns to court and may be resentenced. It is important to keep in perspective the types of offers made in coercive or semi-coercive situations. A patient is unlikely to be offered a *substantially* reduced sentence if they agree to treatment. Moreover, treatment as an alternative is only going to be offered in response to very minor crimes – possession, possibly some small-scale dealing or petty theft. Major crimes, for example, violence committed while intoxicated or linked to drug use, will generally be met with an expectation of retributive justice.

A concern is conflating medical roles (and the involvement of doctors) with those of the penal system; there are also further concerns regarding the effectiveness of coerced therapy in practice, and many of these concerns will depend and indeed transpire or not depending on the individual's perceptions of coercion, and their circumstances and appreciation of the situation. For example, if a person is forced into therapy, they may be more likely to resist treatment or become disparaging towards the medical fraternity (this is a particular concern with pregnant women). This can have significant impact on the individual and those around them and may not achieve treatment goals or exacerbate other clinical and social issues. The degree to which a program is "forced" upon individuals – at the extreme, using constraint – may negatively affect staff attitudes and compromise the treatment quality (Hall 1997). In other circumstances, an individual may be reluctant at first, but will engage and benefit from treatment eventually; in this case, the justification of coercion seems more defensible in terms of consequential utility measurements (that in the end, the individual and society are better off). Of course, the conception of benefit for the individual and the program administrators could well be very different. If a program is grounded in liberty or rights, then it seems that no degree of coercion would be justified in the case of a noncompetent patient.

There are different forms of coerced treatment for drug addiction that vary in the amount of force used (how voluntary the final decision might be) and the options offered and therefore, the degree to which they contravene an individual's freedom and what freedom they have rescinded because of felonious actions. Treatment may also be justified on the basis that the addict has impaired autonomy. This draws on the evidence addicts generally show strong signs of neurobiological change compared to nonaddicts (Leshner 1997; Hyman 2007; Kalant 2010;

Volkow et al. 2010). For example, Caplan (2008) argued that addicted individuals may be temporarily coerced into long-acting drug implant treatments on the grounds of "soft" paternalism – in their best interests – and that this would result in the long-term restoration of the person's autonomy. While, on the one hand, coercion is only as effective as it is applied, and ambivalence to treatment could indicate little intention for long-term abstinence, on the other hand, it is also possible that coerced treatment could restore autonomy – this is one of the arguments for the use of immunological blockade and drug implants approaches that force individuals to undergo a period of compelled abstinence for their own future benefit.

However, it is clear from studies that even those dependent on strongly addictive drugs, such as heroin and nicotine, have the ability to abstain, especially when given incentives. Social incentives, such as family pressure or threat of ending a relationship, are particularly effective (Hasin 1994). It is also clear that individuals generally have the competence to make decisions most of the time despite their drug use. Thus, while autonomy can be impaired in some individuals some of the time, it is not always compromised – for many it is at most perhaps lost temporally or intermittent; and this suggests that forced, or even compelled treatment, remains a human rights issue. Manifestations of brain illness, therefore, should not be taken to be catastrophic to present or future autonomy (Skog 2003). For example, millions of people are ostensibly addicted to nicotine, but that is insufficient grounds to assume they lack the same autonomy and rights as any other person, let alone justify coerced therapy on the basis that they are addicted.

One of the most persuasive arguments for the use of coercion in the treatment of addiction is that it lowers chances of reoffending because punitive-only measures fail to reduce drug use and drug-related crime (Hall 1997; Pedic 1990; Stathis 1991). There is also evidence that injecting drugs in prison increases the transmission of HIV and hepatitis C, suggesting that keeping drug users out of prison in favor of rehabilitation centers would protect them from contracting such diseases (Bell et al. 1992; Hubbard 1989; Ward et al. 1992). Of course, defendants may find the discomfort of a constrained choice more preferable to more serious and further sanctions. Here, the degree of choice becomes critical: because it is difficult to identify the motivation for acceding to treatment (because it is a coerced choice), then the validation of an addict's will become questionable, as do the circumstances that this creates for choices and options. Thus, the benefits of coerced therapy – in terms of a solution to the human and social blight of addiction – within a penal judicial system are debatable (especially if retribution is thought to be involved).

Others might find that the justification for treatment *or* punishment depends on whether the person is willing to commence therapy or not (i.e., willingly take action to stop their criminal activities) and which, in turn, compromises the case for coerced therapy. With already full prisons in some jurisdictions, encouraging individuals to take responsibility and deal with the circumstances which have lead to their (relatively minor) misdemeanors is preferable. In this respect, "treatment" may also be mandated with no other option – such as in "civil commitment programs" – where individuals are remanded into rehabilitation centers for

a minimum period of time. Here, other societal benefits become telling, such as recidivism and reducing the burden of disease. However, such programs are falling out of popularity since they are not as effective as "restrained choice" approaches (Wild et al. 2002).

Whether coercive treatment is ever appropriate remains a long-standing debate in regard to the rights of the individual, the protection of society from drug-associated harm, and the effectiveness of restrained and forced "choices." From whichever viewpoint, these considerations make the coercive use of treatments ethically contentious. At the very least, the safety and effectiveness of these technologies should be established in voluntary individuals before they are trialed in mandated situations.

Concluding Remarks

In this chapter, we have looked at conventional and experimental treatments for addiction. Our analysis proceeded along the lines that there are two predominant models of addiction – the moral and medical models. We conclude by stating that in isolation, both are potentially unhelpful in drug use policies, because, in different ways, they push reactions to compromising extremes. We conclude by offering an alternative to current, predominantly punitive responses to addiction and the neuroscience-based medical model.

Addiction appears to be a transient state and therefore is bereft of a straightforward solution. Punitive responses – the predominant model for current drug policies – are commonly criticized as aggravating by means of declaring a "war" on the circumstances and perpetuation of drug use. In this respect, some bioethicists have encouraged, perhaps unwittingly, the attitude that discouraging problematic drug use requires society to bring down the full force of retributive law upon individuals who flout drug prohibitions. This is a traditionally "hawkish" response to the drug misuse. The medical model, on the other hand, attempts to vindicate the "triumph" of neuroscience over addiction but at the same time can validate heroic and sometimes harsh or overly coercive treatment by the medical fraternity – a questionable role for them to assume. While this debate continues to create polemic camps, the loci of the addicted person have been lost; although they are quite possibly irresponsible and engaging in harmful actions, they could also be desperate for help to break the cycle of use. Perhaps most concerning is that contrary positions lose sight of the circumstances of drug use, which is often linked to conditions of poverty and social instability.

It is important, therefore, that neuroscience informs the addiction debate, that effective treatments – including social interventions and investment in providing opportunities – are considered together, and that whether medications are deployed or not is because they have been shown to reduce the relative harm of addiction, not because their psychopharmacological action is within a category that raises moral conflict, or will require laborious or rowdy political reform. Basing judgments on the premises of moralizing, rather than practical reasoning – that

includes the biological reality of addiction – could result in those who truly seek help being left untreated and diverted unnecessarily into penal systems. This has led to administering suboptimal, even harmful treatments, in order to avoid the use of effective alternative treatments. Effectively treating an individual for problematic drug use could conceivably, therefore, also include addressing social conditions of deprivation.

Throughout this chapter, we have emphasized that a one-size-fits-all approach is unlikely to result in comprehensive solutions to addiction. Perhaps a more fruitful approach would be to channel different people into appropriate treatment programs, which may or may not be linked to other punitive measures. Working out appropriate treatment, therefore, is a task of political will, available resources, and legal and social perceptions of drug use.

Cross-References

▶ Drug Addiction and Criminal Responsibility
▶ Ethics in Neurosurgery
▶ Neuroscience Perspectives on Addiction: Overview
▶ Neurosurgery: Past, Present, and Future
▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues
▶ What Is Addiction Neuroethics?

References

American Psychiatric Association. (1994). *DSM-IV: Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2012). *DSM-V*: *The future of psychiatric diagnosis*. Available at: http://www.dsm5.org/Pages/Default.aspx. Accessed April 2012.

Ashcroft, R., & Franey, C. (2004). Further ethical and social issues in using a cocaine vaccine: Response to Hall and Carter. *Journal of Medical Ethics, 30*(4), 341–343.

Ashcroft, R., Campbell, A., & Capps, B. (2007). Ethical aspects of developments in neuroscience and drug addiction. In D. Nutt, T. Robbins, G. Stimson, M. Ince, & A. Jackson (Eds.), *Drugs and the future: Brain science addiction and society* (pp. 439–465). London: Elsevier.

Australian National Council on Drugs (ANCD). (2012). Position statement: Naltrexone sustained release preparations (injectible & implants). http://www.ancd.org.au/images/PDF/Positionstatements/naltrexonepositionstatement.pdf. Accessed April 2012.

Bell, J., Hall, W., & Byth, K. (1992). Changes in criminal activity after entering methadone maintenance. *British Journal of Addiction, 87*, 251–258.

Caplan, A. (2008). Denying autonomy in order to create it: The paradox of forcing treatment upon addicts. *Addiction, 103*, 1919–1921.

Carter, A., & Hall, W. (2012). *Addiction neuroethics: The promises and perils of neuroscience research on addiction*. London: Cambridge University Press.

Carter, A., & Hall, W. (2011). Proposals to trial deep brain stimulation to treat addiction are premature. *Addiction, 106*(2), 235–237.

Courtwright, D. (2001). *Forces of habit: Drugs and the making of the modern world*. Cambridge, MA: Harvard University Press.

Daniel, M., Martin, A., & Carter, J. (1992). Opiate receptor blockade by naltrexone and mood state after acute physical activity. *British Journal of Sports Medicine, 26*(2), 111–115.

Dean, A. J., Saunders, J. B., Jones, R. T., Young, R. M., Connor, J. P., & Lawford, B. R. (2006). Does naltrexone treatment lead to depression? Findings from a randomized controlled trial in subjects with opioid dependence. *Journal of Psychiatry & Neuroscience, 31*(1), 38–45.

Diefenbach, G. J., Diefenbach, D., Baumeister, A., & West, M. (1999). Portrayal of lobotomy in the popular press: 1935–1960. *Journal of the History of the Neurosciences, 8*(1), 60–69.

Engel, G. (1980). The clinical application of the biopsychosocial model. *The American Journal of Psychiatry, 137*, 535–544.

Gao, G., Wang, X., He, S., Li, W., Wang, Q., Liang, Q., et al. (2003). Clinical study for alleviating opiate drug psychological dependence by a method of ablating the nucleus accumbens with stereotactic surgery. *Stereotactic and Functional Neurosurgery, 81*(1–4), 96–104.

Gartner, C. E., Hall, W. D., Vos, T., Bertram, M. Y., Wallace, A. L., & Lim, S. S. (2007). Assessment of Swedish snus for tobacco harm reduction: An epidemiological modelling study. *Lancet, 369*(9578), 2010–2014.

Hall, W. (1997). The role of legal coercion in the treatment of offenders with alcohol and heroin problems. *Australian and New Zealand Journal of Criminology, 30*, 103–120.

Harwood, H., & Myers, T. (2004). *New treatments for addiction: Behavioural, ethical, legal, and social questions*. Washington, DC: National Academies Press.

Hasin, D. (1994). Treatment/self-help for alcohol-related problems: Relationship to social pressure and alcohol dependence. *Journal of Studies on Alcohol, 55*, 660–666.

Hubbard, R. (1989). *Drug abuse treatment: A national study of effectiveness*. London: University of North Carolina Press.

Hyman, S. (2007). The neurobiology of addiction: Implications for voluntary control of behaviour. *The American Journal of Bioethics, 7*, 8–11.

Kalant, H. (2010). What neurobiology cannot tell us about addiction. *Addiction, 105*, 780–789.

Kass, L. (2003). *Beyond therapy: Biotechnology and the pursuit of happiness*. Washington, DC: President's Council on Bioethics.

Kosten, T., & Owens, S. M. (2005). Immunotherapy for the treatment of drug abuse. *Pharmacology & Therapeutics, 108*(1), 76–85.

Koob, G., & Le Moal, M. (2006). *Neurobiology of addiction*. New York: Academic.

Krack, P., Hariz, M., Baunez, C., Guridi, J., & Obeso, J. (2010). Deep brain stimulation: From neurology to psychiatry? *Trends in Neurosciences, 33*(10), 474–484.

Krupitsky, E., Zvartau, E., & Woody, G. (2010). Use of naltrexone to treat opioid addiction in a country in which methadone and buprenorphine are not available. *Current Psychiatry Reports, 12*(5), 448–453.

Krupitsky, E., & Blokhina, E. A. (2010). Long-acting depot formulations of naltrexone for heroin dependence: A review. *Current Opinion in Psychiatry*. doi:10.1097/YCO.0b013e3283386578.

Leshner, A. (1997). Addiction is a brain disease, and it matters. *Science, 278*, 45–47.

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Ling, W., Casadonte, P., Bigelow, G., Kampman, K. M., Patkar, A., Bailey, G. L., et al. (2010). Buprenorphine implants for treatment of opioid dependence: A randomized controlled trial. *JAMA: The Journal of the American Medical Association, 304*(14), 1576–1583.

Maastricht University Medical Centre (2009). A phase 2B, multi-center, randomized, double-blinded, parallel-arm, study to assess efficacy and safety of 3′-Aminomethylnicotine-P. Aeruginosa r-Exoprotein A conjugate vaccine (NicVAX®) or placebo co-administered with varenicline (Champix®) as an aid in smoking cessation. http://clinicaltrials.gov/ct2/show/NCT00995033. Accessed 6 Oct 2010.

Martell, B., Orson, F., Poling, J., et al. (2009). Cocaine vaccine for the treatment of cocaine dependence in methadone-maintained patients: A randomized, double-blind, placebo-controlled efficacy trial. *Archives of General Psychiatry, 66*, 1116–1123.

Medvedev, S., Anichkov, A., & Poltakov, I. (2003). Physiological mechanisms of the effectiveness of bilateral stereotactic cingulotomy in treatment of strong psychological dependence in drug addiction. *Fiziologiia Cheloveka, 29*, 117–123.

Moreno, A., Azar, M., Warren, N., Dickerson, T., Koob, G., & Janda, K. (2010). A critical evaluation of a nicotine vaccine within a self-administration behavioral model. *Molecular Pharmaceutics, 7*(2), 431–441.

Murphy, M., Checkley, S., Seckl, J., & Lightman, S. (1990). Naloxone inhibits oxytocin release at orgasm in man. *The Journal of Clinical Endocrinology and Metabolism, 71*(4), 1056–1058.

Murray, T. (2004). Ethical issues in immunotherapies or depot medications for substance abuse. In H. J. Harwood & T. G. Myers (Eds.), *New treatments for addiction: Behavioral, ethical legal and social questions* (pp. 188–212). Washington, DC: National Academies Press.

Nutt, D. (2006). Alcohol alternatives: A goal for psychopharmacology? *Journal of Psychopharmacology, 20*(3), 318–320.

Nutt, D., & Lingford-Hughes, A. (2004). Infecting the brain to stop addiction? *Proceedings of the National Academy of Science of the United States of America, 101*(31), 11193–11194.

Orson, F., Kinsey, B., Singh, R., et al. (2008). Substance abuse vaccines. *Annals of the New York Academy of Sciences, 1141*(1), 257–269.

Pedic, F. (1990). *Drug use in prisons: Data collection procedures. A review and recommendations*. Sydney: National Drug and Alcohol Research Centre.

Schlaepfer, T., & Fins, J. (2010). Deep brain stimulation and the neuroethics of responsible publishing: When one is not enough. *JAMA: The Journal of the American Medical Association, 303*, 775–776.

Skog, O. (2003). Addiction: Definitions and mechanisms. In N. Heather & R. E. Vuchinich (Eds.), *Choice, behavioural economics and addiction*. Oxford: Elsevier.

Stathis, H. (1991). Drug use among offenders: A literature review. *Research and Statistics Digit*, NSW Department of Corrective Services.

Volkow, N., Wang, G., Fowler, J., et al. (2010). Addiction: Decreased reward sensitivity and increased expectation sensitivity conspire to overwhelm the brain's control circuit. *BioEssays, 32*, 748–755.

Vocci, F., & Chiang, C. N. (2001). Vaccines against nicotine: how effective are they likely to be in preventing smoking? *CNS Drugs, 15*(7), 505–514.

Ward, J., Mattick, R., & Hall, W. (1992). *Key issues in methadone maintenance treatment*. Kensington: New South Wales University Press.

Ward, J., Mattick, R., & Hall, W. (1998). *Methadone maintenance treatment and other opioid replacement therapies*. Sydney: Harwood Academic Press.

Wild, T., Roberts, A., & Cooper, E. (2002). Compulsory substance abuse treatment: An overview of recent findings and issues. *European Addiction Research, 8*, 84–93.

Wu, H.-M., Wang, X.-L., Chang, C.-W., Li, N., Gao, L., Geng, N., et al. (2010). Preliminary findings in ablating the nucleus accumbens using stereotactic surgery for alleviating psychological dependence on alcohol. *Neuroscience Letters, 473*(2), 77–81.

Yeomans, M., & Gray, R. (2002). Opioid peptides and the control of human ingestive behaviour. *Neuroscience and Biobehavioral Reviews, 26*(6), 713–728.

Drug Addiction and Criminal Responsibility

68

Jeanette Kennett, Nicole A. Vincent, and Anke Snoek

Contents

Abstract

Recent studies reveal some of the neurophysiological mechanisms involved in drug addiction. This prompts some theorists to claim that drug addiction diminishes responsibility. Stephen Morse however rejects this claim. Morse argues that these studies show that drug addiction involves neither compulsion, coercion, nor irrationality. He also adds that addicted people are responsible for becoming addicted and for failing to take measures to manage their addiction. After summarizing relevant neuroscience of addiction literature, this chapter engages critically with Morse to argue that a subgroup of addicted people does meet

J. Kennett (✉) • A. Snoek
Department of Philosophy, Macquarie University, Sydney, NSW, Australia
e-mail: jeanette.kennett@mq.edu.au; anke.snoek@mq.edu.au

N.A. Vincent
Department of Philosophy, Georgia State University, Atlanta, GA, USA

Philosophy Section, Technische Universiteit Delft, Delft, The Netherlands
e-mail: me@nicolevincent.net; nicole.vincent@mq.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_71,

plausible criteria for compulsion, coercion, or irrationality; that few addicted people are fully responsible for becoming addicted; and that some addicted people can be at least partly excused for failing to manage their addiction. Pickard and Lacey's "responsibility without blame" approach is also suggested as a fruitful basis for future work in this field.

Introduction

In recent years, neuroscientific studies have uncovered many of the physical processes and mechanisms involved in drug addiction and addiction-related behaviors. This evidence has prompted some theorists to claim that drug addicts cannot satisfy the conditions for criminal responsibility. If so, how does drug addiction impact on criminal responsibility? Are addicted criminal offenders fully responsible for what they do, or does their addiction somehow diminish their responsibility for the drug-related crimes that they commit?

Stephen Morse has defended the criminal law's commitment to holding people responsible for what they do from similar challenges posed by behavioral genetics and cognitive neuroscience. He has argued that once the conditions of criminal responsibility are properly understood, neuroscientific findings will not support the claim that drug addiction diminishes responsibility. On Morse's account, drug addiction has minimal impact on criminal responsibility because it does not involve compulsion, coercion, or irrationality and because addicted people are responsible for becoming addicted and for failing to take measures to manage their addiction once developed.

After summarizing the most salient points from the neuroscience of addiction literature for the present topic – see ▶ Chaps. 65, "Neuroscience Perspectives on Addiction: Overview" and ▶ 66, "Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response" of this handbook for in-depth discussion of the neuroscientific findings – this chapter engages more closely with Morse's arguments to reveal some of their weaknesses. It concludes that a subgroup of addicted people does meet plausible criteria for compulsion, coercion, or irrationality; that few addicted people are fully responsible for becoming addicted; and that some addicted people can be at least partly excused for failing to manage their addiction.

Drug Addiction and the Brain

Until relatively recently drug addiction was viewed in terms of pleasure seeking and pain avoidance. On this view, people's initial drug use is motivated by desire for pleasure produced by drugs (the drug-induced "high"). Drug tolerance is used to explain why over time increased doses are needed to produce the same pleasurable effects. The explanation offered for why some people do not stop using drugs even when pleasure decreases and the disadvantages of continued use are high – e.g., high financial costs of purchasing ever-increasing quantities of drugs, the drugs'

deleterious health effects, loss of control over use, the risk of a criminal conviction, or even death from overdose – is that they want to avoid unpleasant withdrawal symptoms as well as, to a lesser degree, continue seeking the pleasure experienced during initial drug use (Dalrymple 2006).

The virtue of this way of viewing drug addiction is its simplicity and intuitive plausibility – pleasure seeking and pain avoidance are offered to explain initial experimentation, continued use, and failure to terminate use when costs become high. However, apart from failing to provide a precise account of the mechanisms involved in drug addiction – i.e., what physical changes must occur in order for someone to become addicted to a drug – this account also does not provide a satisfying explanation of the apparent inexplicability of continued drug use once costs of continued use far outweigh any conceivable benefits (see examples in discussion below). Nor does it explain why some people find it so difficult to quit or why people often relapse after long periods of abstinence when the adverse withdrawal symptoms have disappeared.

In recent years neuroscientific studies have supplemented the above picture by suggesting that drug addiction involves complex, gradual, and enduring changes in the brain's reward circuits and control centers. These occur because certain drugs stimulate the release of excessive quantities of the neurotransmitter dopamine. These changes prioritize drug use and sensitize drug users to drug use cues, in effect giving rise to strong and persistent urges to seek out and use these drugs. They also reduce drug users' ability to exercise cognitive control over those urges. Both of these points are elaborated below (Robinson and Berridge 1993; Volkow et al. 2007).

On the first point, the implicated reward circuit – also referred to as the mesolimbic system or mesolimbic pathway – is comprised by the ventral tegmental area, nucleus accumbens, parts of the prefrontal cortex (namely, the orbitofrontal and anterior cingulate cortices), basolateral amygdala, and hippocampus. Animal research shows that this circuit plays a critical role in reinforcement learning by controlling behavior via the association of pleasant outcomes with certain environmental cues and behaviors. For instance, when a person engages in behaviors typically conducive to survival and reproduction – e.g., eating, sex, and positive social interaction – this prompts the release of dopamine into the mesolimbic system from the dopaminergic cells in the ventral tegmental area. This activates the reward circuit producing a pleasurable sensation and creating an association between that behavior and the prospect of reward. From that point onward, environmental cues associated with that behavior are tagged as salient – as carrying the prospect of eliciting pleasure – and thus as worthy of pursuit. However, while this mechanism plays a crucial role in promoting survival – it makes cues for behaviors that are conducive to survival particularly salient, which leads us to notice them and to do things that promote our survival – it is also prone to being "hijacked" by certain drugs (NIDA 2010).

When those drugs are taken, they stimulate the release of dopamine into the reward circuit (or mimic dopamine's action or prevent its reuptake from the synapse), often at levels many times greater than those elicited by normal survival-conducive behaviors. This has several deleterious effects. First, because

levels of dopamine released in response to taking these drugs are so much higher than levels of dopamine released in response to normal survival-conducive behaviors, drug-seeking and drug use and related environmental cues become associated with the promise of greater reward than other behaviors (Robinson and Berridge 1993; Kalivas and Volkow 2005; Volkow et al. 2007). Drug use becomes prioritized over other things, and this, in turn, can lead to repetitive drug use. Second, tolerance for the drug develops with sufficiently frequent use – i.e., greater quantities of the drug are needed to elicit the same degree of pleasure. This is due to a buildup of dopamine in the synapse which blocks subsequent dopamine action and also leads to a decrease ("downregulation") in the number of dopamine receptors at the synapse (Volkow et al. 2007). This effect further decreases the degree of reward elicited by normal survival-conducive behaviors (rather than just de-prioritizing their salience in relation to drug use) and produces a gradual escalation in the dose of drug required to produce the same pleasurable effect. That, in turn, further impairs dopamine's ability to perform its function in the reward circuit. After sufficiently frequent and prolonged use, the person no longer gains pleasure from using the drug but needs to use it merely to feel normal (NIDA 2010). Third, as addiction deepens, adaptations at the molecular level become ingrained in the parts of the brain's reward pathway as connections between neurons involved in coding for the salience of drugs are strengthened (Kasanetz et al. 2010). This results in increased frequency and strength of drug cravings triggered automatically even by casual exposure to cues like the sight of drug paraphernalia.

Research conducted by Nora Volkow and colleagues (Goldstein and Volkow 2002; Kalivas and Volkow 2005; Baler and Volkow 2006; Volkow et al. 2010) suggests that drug use also results in impaired cognitive control. The prefrontal cortex is implicated in cognitive control – it registers different impulses, detects conflicts between them, and selects the impulse that will lead to action. What appears to happen in addiction is that connections between parts of the brain that register reward and salience (the nucleus accumbens and ventral pallidum), memory and learning (the amygdala and hippocampus), and drive and motivation (the orbitofrontal cortex) become hyperactive, while at the same time connections with cognitive control and behavior inhibition (prefrontal cortex) are weakened. This results in a positive feedback loop of the reward-memory-drive aspects of decision-making that bypasses cognitive control. Volkow and colleagues conclude that addiction is thus a disorder of "disrupted self-control" (Goldstein and Volkow 2002; Kalivas and Volkow 2005; Baler and Volkow 2006; Volkow et al. 2010). In effect – according to their interpretation of the neuroscience – the fast impulsive system overtakes the reflective slower system: addiction-related behavior is seen as largely stimulus-driven and subject to minimal conscious control.

In conjunction with studies which suggest that drug addiction is 50 % heritable (Sellman 2010, p. 7; NIDA 2010), this picture of drug addiction has been taken by some as support for the claim that drug addiction is a disease of the brain – albeit an acquired one – and that drug-seeking and drug use are symptoms of a brain disease rather than genuinely intentional actions. Society's focus should therefore shift from holding drug-addicted people responsible for the crimes they commit as

a result of their drug addictions to developing, offering, and maybe even compelling them to undergo medical treatments to cure this disease. Supporters of this position (Leshner 1997; Kosten 1998; Gastfriend 2005; Kasanetz et al. 2010) either explicitly or tacitly endorse the view that once we acquire a greater understanding of the mechanisms and causes involved in drug addiction, all that will be left is to treat drug addiction rather than to punish drug-addicted offenders.

In response to these claims, Stephen Morse has argued that neither the *disease status* of drug addiction nor the *causal role* that implicated brain changes play in subsequent criminal misconduct can diminish addicts' responsibility for the crimes that they commit. Mere causation, on his account, cannot undermine responsibility because if we claimed that causation undermined responsibility, then nobody would be responsible for anything in a physical universe like ours where causation is everywhere (Morse 2006a, p. 405). He also argues that it is not clear why responsibility should be diminished only when behavior is caused by conditions categorized as pathologies (Morse 2011, p. 163), especially when it is far from clear which changes should be labeled as pathological (e.g., see Nadelhoffer and Sinnott-Armstrong 2013 for different accounts of pathology and disease; Vincent 2008 or Watson 2012 on the troubling relationship between pathology/disorder and criminal responsibility). Having rejected these claims about how drug addiction diminishes responsibility for criminal misconduct, Morse then considers how drug addiction fits with more legally conventional accounts of excuse.

Excusing Conditions and Addiction

Could addiction diminish responsibility for crimes committed to support that addiction in some other way though? Ordinary judgments of responsibility rest on a folk-psychological distinction between weakness of will and compulsion (Kennett 2001). The *weak-willed* agent acts in accordance with desires that run contrary to her better judgment; we think that had she exercised self-control, she would have succeeded in bringing her actions into line with her judgment. In contrast, in the case of *compulsion*, we think that the desire or urge on which the agent acts was, at that time, irresistible. Her attempts at self-control fail or would fail. The ordinary view holds that while weak agents are fully responsible for their actions, compelled agents are not fully responsible. But since weak agents and compelled agents may appear indistinguishable from the outside – they may judge and act alike – there may be difficulty in deciding whether the agent performing the act is weak and blameworthy or compelled and at least partly excused. This problem is particularly acute for the criminal law, which raises the question of how courts might determine whether a particular desire was irresistible or merely not resisted? Should courts even try to do this? And could neuroscientific evidence play any role in helping courts to decide?

Standardly it is *compulsion* and *irrationality* – i.e., two possible effects of addiction on people's agency-relevant mental capacities – as well as *coercion* (also called *duress*) that have potential to diminish criminal responsibility. Stephen Morse's view on the availability of these excuses to addicted persons is clear:

> Most addicts should be held responsible for most crimes other than simple purchase or possession, for purely personal use and for use itself. In general addiction does not sufficiently undermine either cognitive or control capacities to excuse the addict for other crimes, and virtually all addicts can be held responsible for not taking the steps to prevent them from engaging in addiction related crime. (2011, p. 173)

Compulsion on Morse's account involves being forced to do something against your will by either an *external* or an *internal* force. For instance, when person A grabs person B's hand and hits person C with it despite B's resistance, that's compulsion by an external factor. Another kind of compulsion occurs when the compelling force comes from inside the agent – for instance, from an irresistible urge which is so strong that it overpowers any efforts at self-control that the agent might make. The thought that some people might be compelled to use drugs is an instance of the second (internal) kind of compulsion.

Genuine compulsion, according to both those who endorse a *choice model* of addiction (e.g., Heyman 2009; Pickard 2011; Morse 2011) and those who endorse a *brain disease model* (Leshner 1997; Volkow et al. 2010), bypasses intention and robs behavior of its status as action. But choice theorists argue, against proponents of the disease model, that drug use cannot count as compelled because it is not like a tic or a reflex. It is not, Morse insists, analogous to hanging by one's fingernails over a cliff where there is a physical limit on one's powers of resistance. He argues that if addicted people's urges to obtain and use drugs were literally irresistible, then no addict should ever be able to resist using drugs in response to incentives. But yet, as choice theorists point out, drug use is responsive to local contingencies such as price or the presence of the police, which suggests that it is intentional. Drug users can certainly *delay* using if the stakes are sufficiently high, for instance as in Morse's example of a person facing court for public drunkenness who restricted himself to a single drink on the morning of his court appearance. Heyman (2009) further argues that everyone *stops* using drugs when the costs become too high, citing data that most people who have ever satisfied the diagnostic criteria for substance dependence eventually quit without treatment. In fact, the data he cites indicate that drug users are capable of exercising control over their use given even trivial incentives. He takes this to indicate that the choice to use drugs and any associated criminal conduct is not compelled. It is accordingly voluntary behavior for which the user is properly held responsible.

In contrast, the strength model of self-control described by Baumeister (2003) sees self-control as a depletable resource. Self-control, or strength of will, requires cognitive resources and these are limited (Levy 2011). On this model, a person who is addicted may initially resist the urge to use drugs, but eventually, if the urge persists, and without any other interventions, ego-depletion occurs, their capacities for resistance are exhausted, and they succumb much like the cliff-hanger. The strength model can thus explain the facts upon which the choice theorists rely while being consistent with the view that for some defendants persistent cravings over time overwhelm their powers of resistance. But even here choice theorists argue that giving in to urges to use drugs is not akin to losing one's grip. Morse argues that because giving in involves intentional action, "it is [thus] distinguishable from purely mechanical signs and symptoms" and so compulsion is absent (Morse 2011, p. 173).

At issue is precisely *how should compulsion be characterized*? If, by definition, nothing that counts as *action* can be compelled, then the defense of compulsion will only be available in situations of external force. After all, if one's internal states mechanistically produce behavior which is not action, (because not intentional) then the relevant defense would surely be one of automatism. One must however concede that the defense of automatism is unlikely to be available for crimes committed in the course of addiction. But if that is right, then is it not plausible that both *choice* and *disease* theorists might have the wrong account of compulsion? That they set the bar too high and mischaracterize the addict's predicament? Here is an alternative characterization of compulsion which does not yield the result that the behavior at issue ends up being viewed as non-intentional. In addiction, persistent and insistent drug-related thoughts and cravings dominate the agent's attention and eventually exhaust their self-control resources. At some point the agent loses their grip (metaphorically speaking) *on their prior resolution or intention to abstain* (see Holton 2009 and Levy 2006 on judgment shift in the face of temptation), and then they use drugs intentionally. Kennett (2001, 2013) proposes that the correct analysis of the ordinary notion of compulsive motivation is of motivation that is largely impervious both to the agent's *values* and to available techniques of self-control, and there is good reason to believe that this condition is often satisfied in addiction.

However, even if the mechanistic account of compulsion were correct and even if addicts were not like the cliff-hanger, it still would not follow that addicts are merely weak-willed. Addicts might still have an excuse for drug-related crime, if it could be shown that drug cravings are *coercive* in the sense that failure to act on them would exact a cost on them which it is unreasonable to expect them to bear. In law, Morse tells us, duress obtains "if the defendant is threatened with the use of deadly force or grievous bodily harm . . .unless the defendant commits an equally or more serious crime" (Morse 2006b, p. 71). The excuse of duress thus does not impugn the agent's rationality or capacities for self-control. Instead it pictures the agent as facing a particularly hard choice. The law uses a "moralized objective standard that uses the person of reasonable firmness as the criterion" of what it is reasonable or not reasonable to resist (Morse 2011, p. 185). Threats of death or grievous bodily harm are objective indicators of a hard-choice situation. It would be unfair of the law to expect someone to resist in such circumstances, even if resistance were possible. For instance, when a bank robber threatens "The money or your life," the teller is coerced into handing over the money. Given the awful situation that she was in – a choice of sacrificing her life or handing over a substantial amount of the bank's money – we excuse the teller for doing what would otherwise have been wrong.

A central feature of coercion is that it involves hard choices. If that is not part of the scenario, then we do not have coercion. A bank teller who yields to the threat "The money or I'll call you a rude name" would not have recourse to the defense of coercion. Morse argues that in the case of criminal activity to support drug use, the choices that addicted people face are more like the second bank teller's choices. He denies that the pain of withdrawal or the prospect of continued anhedonia is sufficient for addicts' actions to fall under the excuse of coercion or duress. In most cases, he says, withdrawal "is not terribly painful" and can be medically managed.

While addicts may (rationally) take drugs to avoid dysphoria, they do not fear it in the way that death is feared. This conclusion appears to follow from his prior argument that urges to use drugs are not compulsive; if the addict were threatened with instant death if he seeks or uses drugs, then assuming he wants to live "as much as the cliff-hanger does, no addict would yield to the desire" (Morse 2011, p. 180). The cliff-hanger, by contrast, will eventually fall no matter what. Morse presumably thinks the same would be true if an addicted user were threatened with grievous bodily harm. If the addict would (rationally) choose dysphoria over death or grievous bodily harm, then the threat of dysphoria does not constitute a hard choice in the required legal sense to establish duress. He further suggests that because dysphoria and craving are "more purely subjective than death or grievous bodily injury," they are more difficult to assess than fear of death or grievous bodily injury. Few addicts he thinks "would succeed with a hard-choice excuse."

There are however at least two problems with Morse's blanket dismissal of coercion as an excuse in addiction. First, it is not clear that fear of death or grievous bodily harm should be the standard required for duress in minor property crimes, where no person is threatened or harmed (say breaking into a parked car and taking the wallet lying in full view on the seat or stealing the coins from a pay phone), or for low level dealing to other drug users. The threat or continuation of severe dysphoria might be sufficiently coercive in such circumstances to provide at least some excuse for the offence, just as severe hunger might. Second, it is not true of all addicts that the threat of death or grievous bodily harm deters them from using or that compelling behavioral evidence of this cannot be obtained. To the contrary, there are numerous accounts of hard-core addicts who continue using while extremely ill, who know that death is a very probable outcome, and who do so without wishing to die. In an ongoing study being conducted by two of the authors, respondents showed a keen awareness of the very real risk of death from their continued drug use, as well as serious health consequences, amounting to grievous bodily harms that they suffered daily:

> Oh it's very harmful, it's almost killed me, I had cancer last year [. . .], I had alcoholic hepatitis, I was in the hospital almost five times and I had cellulitis four times . . . yeah it really hasn't done me any favors. (R10)
>
> I think now I have kidney or liver problems, [. . .] because certain drinks. . . the pain is just . . . in my lower back, the kidney area [. . .] I can't stand up, I have to lay down and it . . . yeah, like I literally can't get up, it passes after about 45 minutes I think, I think it's just while the kidneys process it or something and it passes, but it's really intense pain. . . (R4)
>
> All my family died from grog, all my brothers, sisters, my mum . . . I found [. . .] my mum dead on the floor from grog. (R13)
>
> [A]nother thing that's affected me is so many of my friends have died of drug overdose, like my first friend killed herself in 1997 and since then I've been to like 30 funerals, 40 maybe. (R53)
>
> [T]here's 12 of us started out together. And I think two are in institutions and the rest are dead, and I'm here. So I guess, yeah, I'm counting my blessings there. Any time that could have been me. (R39)
>
> You've always got that in the back of your mind you're going to have a seizure, or you may not wake up. (R25)

These examples demonstrate that at least some addicted individuals do not cease using even when faced with imminent death or grievous bodily harm. Some of them – those who are addicted to expensive illicit drugs – will commit crimes to support their use. Does this mean that they fear dysphoria more than death? This is not obvious. On Morse's account of coercion, coerced choices are difficult but not baffling in the way that our respondents' choices to continue drug use in the face of death surely are. Is their situation best described as one of coercion, compulsion, extreme irrationality, or simply as a personal choice for which they are responsible in the ordinary way?

Morse argues that if addiction excuses criminal conduct, it must hinge on the addict's *irrationality*. But in what sense are these users irrational? They do not have any unwarranted false beliefs about the likely consequences of their activities. The view advanced here is that the desires of the addicted person are irrational in that they are not responsive in the ordinary way either to the person's own *beliefs* about the likely outcomes of their use or to their *values*. Deliberation for such persons cannot get a grip on choice. Indeed, many of the respondents in the above study expressed deep sadness at their failure to live the kind of life they would have valued living and the loss of a home, family, and career opportunities to their drug use. Here it is plausible that the neuroscience of addiction *can* shed light on this puzzling phenomenon. According to Berridge:

> Human drug addiction may be a special illustration of intense "wanting" that results from permanent sensitization of mesocorticolimbic systems. Sensitized "wanting" may rise to quite irrational levels. That is, the intensity of cue-triggered "wanting" to take drugs for brain-sensitized addicts could outstrip their "liking" even for pleasant drugs, outstrip their expectation of how much they will like the drugs, and outlast any feelings of withdrawal if they stop. Brain-sensitized addicts may be unable to give a reason for their drug taking in such a case. Indeed, there is no reason, there is only a cause for why they "want" so much. (Berridge 2009, pp. 384–385, internal citations removed)

In line with this, respondents in the above study of users would generate reasons for use in which escape from dysphoria was sometimes mentioned as was pleasure. But for many these reasons for use fell by the wayside in the course of their addiction:

> I used it as a coping mechanism sort of thing. Yeah but now it's just . . . it's not even fun anymore really, it just sort of becomes a . . . I don't know, more or less like a chore I suppose but yeah I just . . . I want to get away from it. (R29)
>
> It's . . . there was reason, early part, until I came to understand why I was behaving the way I was behaving. So in . . . no, not now. No. There's no reason. (R39)
>
> When I was 20, 30 when I was 40 my drinking was good, I had good times on the drink,. . . I'm just drinking for nothing [. . .] I'm just drinking for drinking sake now. (R24)

For these respondents the desire or urge to use drugs is largely detached from their assessment of the consequences of such use, or their desire for the effects that the drugs once used to produce, and from their overarching values. Their subjective experiences accord with Berridge's claim that in addiction wanting and liking become dissociated and with the neurological evidence he produces. Now, as Morse duly acknowledges, neurobiological and psychological states that are signs

of addiction are states that “are mechanistically produced.” The question is how these states might affect responsibility. Insofar as the addicted person acts “for nothing,” for “no reason” that they can fathom, it appears that their defense may be *irrationality* rather than *coercion*. Insofar as they act unwillingly, doing something they neither like nor value, it appears that they might satisfy the ordinary, if not the legal, understanding of *duress* or *compulsion*. They do not act of their own free will.

On any of these interpretations, it would be unjust to hold such individuals, in whom wanting and liking have become dissociated, fully responsible for criminal actions arising from their addiction unless the following arguments succeed. Addicts must be (1) responsible for becoming addicts *and* this responsibility for actions in the past at a time when many of them were adolescents or children is sufficient to support full criminal responsibility as adults, or (2) the coercion, compulsion, or irrationality to which they are subject does not prevent the exercise of diachronic self-control and the availability of diachronic self-control is sufficient to support a finding of criminal responsibility. The next section contends that there are hard cases in which neither of these conditions are met.

Responsibility for Becoming Addicted

Morse argues that addicted people have only themselves to blame for their addiction. In his view, apart from infants born to mothers who used drugs during pregnancy, nobody is compelled or coerced into drug use. Rather, initial experimentation with drugs is usually a voluntary choice. People normally start experimenting with drugs in mid- to late adolescence, by which stage they are sufficiently psychologically mature to be held responsible in a court of law. And the risks associated with drug use are well publicized so there is little excuse for not being aware and taking heed of them. Bad social circumstances are likewise not a ground for excuse since few people from underprivileged backgrounds turn to crime. And to the extent that addicted people are to blame for their own predicament, they are not entitled to cite their predicament as an excuse.

While it is possible that a proportion of addicted people bear at least some responsibility for their addiction, a number of considerations are presented below to explain why such responsibility may in fact be absent or diluted to such a degree that it is highly dubious that this could be sufficient to ground a finding of criminal responsibility at a much later date and to extinguish excuses such as compulsion or coercion which might then be present.

First, a substantial portion of addicted people are first exposed to drugs at an extremely early age. One of the users that was interviewed in the above study received his first line of speed from his older brother’s friend at the age of ten. Others grew up with alcoholic parents who treated them as “drinking buddies” from a very young age. Another subject reports:

> I can't remember before I used. It's my whole life. Drugs have been my whole life. I remember at five sitting there drinking wine and eating mull cookies, so that's all I really remember is drug use. (R38)

Second, from a developmental perspective, adolescence is an especially vulnerable period. Smith, Xiao, and Bechara (2012) as well as Crone, Vedel, and van der Molen (2003) report that although most mental capacities gradually and steadily improve from childhood into adulthood, not all mental capacities follow this linear developmental trajectory. For instance, it turns out that younger and more developmentally immature children perform better on the Iowa Gambling Task than older and developmentally more mature early adolescents, not until later adolescence does performance recover. Smith and colleagues explain that:

> [t]his trajectory is thought to coincide with asymmetric neural development in early adolescents, with relatively overactive striatal regions creating impulsive reward-driven responses [in adolescence] that may go unchecked by the slower developing inhibitive frontal cortex. (2012, p. 1180)

This suggests that at some points during the process of maturation, mental capacity differences between adults and adolescents can be significant and that the reflection and foresight expected of adults is not an appropriate standard to apply to adolescents.

Third, the implicit focus on illegal drugs that permeates this discussion is regrettable. It makes it seem as if people must always surely be blameworthy for their initial decision to use potentially addictive substances. But this ignores the fact that most people in Western societies use addictive drugs recreationally at some stage in their lives, and most people do not become addicted. Addiction occurs at a fairly stable rate among exposed individuals, and those who become addicted may have taken no more risks than those who do not, but rather they were simply the unlucky ones who have a genetic or other predisposition towards becoming addicted (Sellman 2010, p. 7; NIDA 2010). Because the illegality of drugs is imperfectly correlated with their addictiveness (O'Brien and McLellan 1996), those who use and subsequently become addicted to illegal drugs might at best be blameworthy for breaking the law but not for taking unreasonable risks of addiction. In any case, as examples quoted from the above study illustrate, illegal drug users are more likely to be exposed to drug use at a very early age when they could not be blamed either for breaking the law or for taking unreasonable risks. Further, the focus on illegal drugs overlooks the promotional activities of companies who manufacture and market tobacco and alcohol, as well as governments who sanction this use and collect taxation revenue from sales of these substances. These groups are arguably also at least partly responsible for many people's addictions.

Finally, there is disquiet in philosophy about whether this "historical tracing" (Fischer and Ravizza 1998) approach can justify the practice of discounting the exculpatory value of mental incapacities for which we are responsible. One problem is that we must often go back so far in time to find the initial faulty action – the one which allegedly started a person off on the path of eventually committing a criminal offence as a result of being addicted. It stretches credulity to suggest that

the reason why this person is *now* blameworthy, despite their currently impaired mental capacities, is because their much younger self should have foreseen this eventual outcome all those years ago. Put simply, nobody (particularly children and adolescents) has nor can be expected to have that much foresight.

Of course it might be argued that the person whose drug use becomes problematic will experience negative outcomes well in advance of contact with the legal system. They will have had both ample warning and ample opportunity to change their ways. The problem with using this point to dismiss or to depreciate the exculpatory value of addiction is that gradual changes – such as increasing tolerance – may pass unnoticed, and by the time the negative consequences are forced upon the person's attention, their addiction is already established. Given that addiction may be established at a relatively young age, the point about foresight still holds.

The other problem with the tracing approach is that the *content* for which we wish to blame people, i.e., the criminal offence with which they are charged (e.g., possession or use of drugs, burglary, mugging, or perhaps even killing someone), ends up being very different to the content of the allegedly faulty action to which we trace things back (e.g., that they experimented with a potentially addictive substance) (Benchimol 2011).

These four considerations entail that responsibility for becoming addicted should not be a default assumption in criminal proceedings. Moreover, even when the defendant does bear some responsibility for becoming addicted, this will usually be an insufficient basis upon which to assign responsibility for their later criminal conduct.

Diachronic Self-Control and Responsibility in Addiction

Diachronic self-control is a form of control exercised in advance of an experienced temptation. It has two broad forms (Kennett 2001). First, the agent may manipulate her future circumstances to minimize the chance of the temptation arising or to ensure that the temptation is not acted upon. Take the case of smoking. Someone who is trying to quit smoking may decide not to go to Friday night drinks at the pub or to put herself in other situations known to trigger cravings. Or she might raise the cost of satisfying the desire for a cigarette by making a public bet or commitment to remain abstinent. Or she might leave her wallet at home so that she can't buy cigarettes on the way home from work when the craving usually arises. Second, if these strategies don't work, she may try to act directly on the desire to smoke by, say, taking a medication that reduces cravings, or by engaging in psychotherapy.

The advantage of diachronic strategies of self-control is that they don't require as much willpower as synchronic forms of control since they can be put into place when we are not under the distorting or overpowering influences of temptation. The addicted person can at those times recognize the reasons they have for quitting and take steps towards this end. The thought is that the predictable craving for drugs in

the evening does not prevent one from seeing a counselor or a doctor in the morning when those cravings are absent or muted. Running with this line of thought, Stephen Morse argues:

> Even if addicts are sufficiently *irrational* or are so "internally *coerced*" as to warrant mitigation or excuse at the time they commit their substance related crimes, they should still be held responsible [. . .] because *they have lucid, rational intervals between episodes of use during which they could* act on the good reasons to *seek help quitting or* otherwise to *take steps to avoid engaging in harmful drug-related behaviour*. (191, emphasis added)

Morse's argument is persuasive. The availability of techniques of diachronic self-control casts the net of responsibility wide and makes it more difficult to establish an excuse to criminal conduct on grounds of internal coercion or diminished rationality. It is probable that many people charged with drug-related crime have not faced a sufficiently hard choice to count as being internally compelled, or they have had adequate opportunity to exercise diachronic self-control with respect to their offending behavior. Nevertheless, Morse overlooks a category of users – a clinical subgroup – to whom this argument does not apply, and it is argued below that holding *this* category of addicts fully responsible for their criminal misconduct may well be unjust.

Limitations of Diachronic Self-Control in Addiction

There are two kinds of barriers to successful diachronic control in hard cases of addiction – that is, those cases that fit the diagnostic criteria of being chronic and relapsing, which persist beyond the late 1920s, and that involve clear and significant negative health and personal consequences. *Internal* barriers include (1) myopia for the future, an incapacity to foresee or pay attention to future consequences; (2) an inability to project oneself into, and so to plan realistically for, a future in which one is abstinent; (3) poor physical health; and (4) monopolization of attention which restricts one's thinking to an almost wholly synchronic focus. *External* barriers include (1) a lack or nonavailability of treatment options, (2) ineffective treatments, and (3) environmental factors that a person can't control, such as proliferation of liquor outlets and intrusive advertising, or a culture of drug use that they can't easily remove themselves from for legitimate reasons like poverty. Two of these factors which have particularly pronounced effects on this clinical subgroup's ability to exercise diachronic self-control are elaborated upon below.

An important aspect of diachronic self-control is mental time travel, i.e., the ability to project oneself into a personal future. Mental time travel allows us to plan for the future, secure our projects, and unify ourselves across time. It underwrites exercises of diachronic self-control. If hard-core addicts are highly fragmented or synchronic selves as some of the data suggests (e.g., Ainslie 2011), then they will not be able to project themselves into a personal future or identify with their future self in ways which motivate successful self-control. The temporary absence of craving is not sufficient to secure the availability of strategies of diachronic self-control in addiction.

Neurological and behavioral evidence suggests that diachronic capacities can be impaired in addiction. The work of Ainslie and colleagues shows that addicts apply a steeper discount rate to future consequences than others, suggesting that these consequences are less vivid to them. Other work done by Bechara, Dolan, and Hindes (2002) and Bechara (2005) on the Iowa Gambling Task also suggests that there is a subcategory of users who are insensitive to future consequences. This task registers on the one hand a choice for low- or high-risk decks, and on the other hand via skin conductance response registers emotional arousal and stress. Normal participants initially choose cards at random and over time develop a preference for the low-risk decks that give lower rewards but pay more in the long run. Before they can consciously describe why they have chosen the low-risk decks, they develop an anticipatory skin conductance response before choosing from the high-risk deck, indicating a feeling of stress. Persons with damage to the prefrontal cortex do not develop an anticipatory skin conductance response and keep choosing cards from the high-risk deck. Bechara distinguished three subgroups when he presented the test to substance-dependent people. The first group reacted like the normal participants; they developed the skin conductance response and started choosing the low-risk decks. This group was also higher functioning than the other two groups. The second subgroup reacted in the same way as the participants with prefrontal cortex damage: they didn't show anticipatory skin conductance responses and kept choosing from the high-risk decks. Bechara suggests that this group is "myopic for the future." The third subgroup kept choosing the high-risk decks, but they showed high skin conductance anticipatory responses (excitement) every time they chose a card from the high-risk decks. Bechara suggests that this group is highly sensitive to reward with poor impulse control.

Bechara's work has the following significance. Drug users in the first group who are responsive to incentives in the usual way would probably not satisfy the conditions for an excuse of coercion or compulsion or irrationality. One might suspect this group encompasses the majority of users who, as Heyman points out, cease using drugs without assistance as they mature and acquire life responsibilities. The third group lacks *synchronic* self-control. They may be able to institute some measures of diachronic self-control, but given their hypersensitivity to reward their self-control will be very vulnerable to external undermining by drug cues and the like. The second group shows a reduced capacity to imagine future consequences and to learn from negative consequences. This group cannot readily avail themselves of the techniques of diachronic self-control. Depending on the severity of their impairment, their responsibility for offending behavior may also be reduced.

The above use of the Iowa test or related neuroimaging in the legal context to determine which subgroup an addicted offender belongs to, or to assess their capacities for self-control, is only speculative at this stage. As Glicksohn et al. (2007, p. 202) point out, a sizeable minority of normal subjects also show impaired performance on the Gambling Task. However, if more specific and sensitive tests became available, they could, in principle, be used to help discriminate between those people who *cannot* control their drug use without considerable external

supports and targeted treatment, and those people who simply *do not*. The above should certainly not be taken to suggest that such technologies could by themselves determine whether responsibility was diminished or not. But it is possible that neuroscientific evidence could supplement behavioral and clinical evidence and, importantly, could be used at the sentencing stage to suggest the response to the offender that would most likely be effective in addressing offending behavior – whether pharmacological treatment, behavior modification therapies, or punishment.

Another important way in which addicts can exercise diachronic self-control is via seeking treatment for their addiction. Even if addiction is a disease, Bonnie (2002) points out that having a disease is not of itself an excusing condition. With this status comes responsibility – responsibility for seeking and complying with treatment. It is true that in addiction the rates of seeking treatment voluntarily and complying with it tend to be low (Chandler et al. 2009, p. 189). However, although treatment-seeking is *delayed* in addiction, most long-term addicts do seek treatment numerous times. Emmelkamp and Vedel (2006) estimate that on average it takes people 8 years and many tries for individuals to successfully overcome their addiction. The problem, though, is that at present there are actually no gold standard treatments available for addiction. Seeking treatment is often further complicated by waiting lists, costs, and exclusion criteria (e.g., for the use of methadone or psychiatric comorbidity). The lack of availability of effective treatment and support for those with serious addictions, particularly for the poor and disadvantaged, constitutes a serious external barrier to the successful exercise of diachronic self-control for which the addicted person cannot be held responsible. After all, there can only be a responsibility to do what is in one's power – i.e., to exercise the strategies of diachronic self-control that *are* available and effective – but if none are available or effective, then this is hardly something for which an addicted offender can be blamed.

Conclusions and Future Directions: Responsibility Without Blame

It is likely that many, perhaps most, people who satisfy the criteria for dependence on either licit or illicit drugs will lack a legal excuse for criminal offences committed pursuant to their drug use. Their drug use is responsive to a sufficient range of incentives, and while they may find it difficult to refrain from using drugs and this may involve some discomfort, this will not likely rise to a level that would provide them with a complete or partial excuse for drug-related criminal offences. For such people the criminal law and threat of punishment is capable of acting as a deterrent as well as a reminder of their responsibilities to develop and to exercise self-control.

However, attempts to refute or dismiss the neuroscientific evidence of responsibility-relevant impairments in hard-core users, or to draw conclusions about the responsibility of all addicted people by reference to the capacities of the majority of substance-dependent people, relies on a faulty generalization. It assumes that addiction is homogenous so that evidence that *most* addicted persons are responsive to incentives and can exercise control is taken as evidence that *all* can. It has been

argued that for a significant subgroup of substance-dependent people – those hard-core long-term addicts for whom numerous attempts to quit have failed – the issue of responsibility is not settled by reference to what most addicts can do. This is the group that the above discussion has focused upon and for whom the cited scientific evidence seems particularly relevant.

In regard to these people, it is hard to resist the conclusion that they are relevantly compelled, coerced, or irrational *and* that they have either not had adequate opportunity to instigate self-control strategies, including receiving treatment that would have prevented their offending, *or* that they have mental impairments that render those strategies largely unavailable or ineffective. It seems unjust to hold these individuals fully responsible for their misconduct and to blame them for it. Moreover, given that a central criterion for addiction is continued use in the face of significant negative consequences and given that some addicts continue using despite the daily threat of death and the actuality of pain and serious bodily harm, the threat or the actuality of legal punishment is unlikely to have any further disincentive effect. It only inflicts further misery and disadvantage upon this group.

However, the idea that addiction could provide a substantial or complete defense to serious criminal misconduct and that addicts whose criminal offences have harmed others should simply be released onto the streets is also unpalatable. This raises the question of what dispositions ought to be available to courts. Although the constraints of space do not permit a detailed elaboration of the following point, there is much that commends itself in Nicola Lacey and Hannah Pickard's (2012) recent "responsibility without blame" approach. On Lacey and Pickard's account, the only effective way to undo the impairments to agency that spring from drug addiction is via employing the fragments of agency that addicted people still retain. Lacey and Pickard suggest that by treating people as responsible agents and by making them accountable for their decisions, we can bootstrap their agency and responsibility into existence. However, blame is counterproductive to this therapeutic enterprise on their account since it further undermines the individual's sense of agency by reinforcing feelings of failure and inadequacy. And for this reason they suggest that an approach which has been successfully employed by clinicians – that of responsibility without blame – could also be usefully carried across into the legal context. This is, admittedly, a forward-looking approach to responsibility, in contrast to the backward-looking approach that is currently the basis for retributive criminal punishment practices. However, despite being forward-looking, it is not simply a consequentialist approach, because it does not give up on the core notions of agency and responsibility in favor of a system of incentives and disincentives. Rather, it has at its core the Kantian notion of respect for persons, which it aims to promote and to foster by treating people as responsible for their misconduct (without at the same time blaming them for it) in order to help them to develop the mental capacities that agents need to possess. Arguably this is what specialized drug courts attempt to do, and the broader implications of Lacey and Pickard's approach for the law deserve more detailed consideration.

Acknowledgments The authors are grateful to the Australian Research Council for their support of project DP 1094144 *Addiction, Moral Identity and Moral Agency: Integrating Theoretical and Empirical Approaches*. They especially thank the participants and staff of the Drug and Alcohol Services unit of St Vincent's Hospital (Sydney) – in particular, Rankin Court and Gorman House – as well as Wayne Hall and Adrian Carter for their helpful comments and practical editorial on this piece.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Determinism and Its Relevance to the Free-Will Question
▶ Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response
▶ Ethical Issues in the Treatment of Addiction
▶ Free Will and Experimental Philosophy: An Intervention
▶ Free Will, Agency, and the Cultural, Reflexive Brain
▶ Moral Cognition: Introduction
▶ Neurolaw: Introduction
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play
▶ Neuroscience Perspectives on Addiction: Overview
▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues
▶ What Is Addiction Neuroethics?

References

Ainslie, G. (2011). Free will as recursive self-prediction: Does a deterministic mechanism reduce responsibility. In J. Poland & G. Graham (Eds.), *Addiction and responsibility* (pp. 55–88). Cambridge, MA: MIT.

Baler, R. D., & Volkow, N. D. (2006). Drug addiction: The neurobiology of disrupted self-control. *Trends in Molecular Medicine, 12*(12), 559–566.

Baumeister, R. F. (2003). Ego depletion and self-regulation failure: A resource model of self-control. *Alcoholism-Clinical and Experimental Research, 27*(2), 281–284.

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. *Nature Neuroscience, 8*(11), 1458–1463.

Bechara, A., Dolan, S., & Hindes, A. (2002). Decision-making and addiction (part II): Myopia for the future or hypersensitivity to reward? *Neuropsychologia, 40*(10), 1690–1705.

Benchimol, J. (2011). The moral significance of unintentional omission: Comparing will-centered and non-will-centered accounts of moral responsibility. In N. Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral responsibility: Beyond free will and determinism* (pp. 101–120). Dordrecht: Springer.

Berridge, K. C. (2009). Wanting and liking: Observations from the neuroscience and psychology laboratory. *Inquiry, 52*(4), 378–398.

Bonnie, R. J. (2002). Responsibility for addiction. *The Journal of the American Academy of Psychiatry and the Law, 30*(2), 405–413.

Chandler, R. K., Fletcher, B. W., & Volkow, N. D. (2009). Treating drug abuse and addiction in the criminal justice system: Improving public health and safety. *JAMA: The Journal of the American Medical Association, 301*(2), 183–190.

Crone, E. A., Vendel, I., & van der Molen, M. W. (2003). Decision-making in disinhibited adolescents and adults: Insensitivity to future consequences or driven by immediate reward? *Personality and Individual Differences, 35*(7), 1625–1641.

Dalrymple, T. (2006). *Romancing opiates: Pharmacological lies and the addiction bureaucracy* (1st ed.). New York: Encounter Books.

Emmelkamp, P. M. G., & Vedel, E. (2006). *Evidence-based treatment for alcohol and drug abuse: A practitioner's guide to theory, methods, and practice*. London: Routledge.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge, MA/New York: Cambridge University Press.

Gastfriend, D. R. (2005). Physician substance abuse and recovery: What does it mean for physicians – and everyone else. *Journal of the American Medical Association, 293*(12), 1513–1515.

Glicksohn, J., Naor-Ziv, R., & Leshem, R. (2007). Impulsive decision making: Learning to gamble wisely? *Cognition, 105*, 195–205.

Goldstein, R. Z., & Volkow, N. D. (2002). Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *The American Journal of Psychiatry, 159*(10), 1642–1652.

Heyman, G. M. (2009). *Addiction: A disorder of choice*. Cambridge, MA: Harvard University Press.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford: Oxford University Press.

Kalivas, P. W., & Volkow, N. D. (2005). The neural basis of addiction: A pathology of motivation and choice. *The American Journal of Psychiatry, 162*(8), 1403–1413.

Kasanetz, F., Deroche-Gamonet, V., Berson, N., Balado, E., Lafourcade, M., Manzoni, O., & Piazza, P. V. (2010). Transition to addiction is associated with a persistent impairment in synaptic plasticity. *Science, 328*(25), 1709–1712.

Kennett, J. (2001). *Agency and responsibility: A common-sense moral psychology*. Oxford/New York: Clarendon Press/Oxford University Press.

Kennett, J. (2013). Just say no? Addiction and the elements of self-control. In N. Levy (Ed.), *Addiction and self –control*. Oxford: Oxford University Press.

Kosten, T. (1998). Addiction as a brain disease. *The American Journal of Psychiatry, 155*(6), 711–713.

Lacey, N., & Pickard, H. (2012). From the consulting room to the court room? Taking the clinical model of responsibility without blame into the legal realm. *Oxford Journal of Legal Studies*, 1–29, http://ojls.oxfordjournals.org/content/early/2012/11/19/ojls.gqs028.full

Leshner, A. I. (1997). Addiction is a brain disease, and it matters. *Science, 278*(3), 45–47.

Levy, N. (2006). Autonomy and addiction. *Canadian Journal of Philosophy, 36*(3), 427–447.

Levy, N. (2011). Addiction, responsibility and ego-depletion. In J. Poland & G. Graham (Eds.), *Addiction and responsibility*. Cambridge, MA: MIT Press.

Morse, S. J. (2006a). Brain overclaim syndrome and criminal responsibility: A diagnostic note. *Ohio State Journal of Criminal Law, 3*, 397–412.

Morse, S. J. (2006b). Addiction, genetics, and criminal responsibility. *Law and Contemporary Problems, 69*, 165–207.

Morse, S. J. (2011). Addiction and criminal responsibility. In J. Poland & G. Graham (Eds.), *Addiction and responsibility* (pp. 159–199). New York: The MIT Press.

Nadelhoffer, T., & Sinnott-Armstrong, W. (2013). Is psychopathy a mental disorder? In N. Vincent (Ed.), *Neuroscience and legal responsibility* (pp. 227–253). New York: Oxford University Press.

NIDA. (2010). Drugs, brains and behavior: The science of addiction. National Institutes of Health, U.S. Department of Health and Human Services. Accessed on 20130402 from http://www.drugabuse.gov/sites/default/files/sciofaddiction.pdf and http://www.drugabuse.gov/publications/science-addiction/drugs-brain

O'Brien, C. P., & McLellan, A. T. (1996). Myths about the treatment of addiction. *Lancet, 347*(8996), 237–240.

Pickard, H. (2011). Responsibility without blame, empathy and the effective treatment of personality disorder. *Philosophy, Psychiatry and Psychology, 18*(3), 209–224.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews, 18*(3), 247–291.

Sellman, D. (2010). The 10 most important things known about addiction. *Addiction, 105*(1), 6–13.

Smith, D. G., Xiao, L., & Bechara, A. (2012). Decision making in children and adolescents: Impaired Iowa gambling task performance in early adolescence. *Developmental Psychology, 48*(4), 1180–1187.

Vincent, N. (2008). Responsibility, dysfunction and capacity. *Neuroethics, 1*(3), 199–204.

Volkow, N. D., Fowler, J. S., Wang, G. J., Swanson, J. M., & Telang, F. (2007). Dopamine in drug abuse and addiction: Results of imaging studies and treatment implications. *Archives of Neurology, 64*(11), 1575–1579.

Volkow, N. D., Fowler, J. S., Wang, G. J., Telang, F., Logan, J., Jayne, M., et al. (2010). Cognitive control of drug craving inhibits brain reward regions in cocaine abusers. *NeuroImage, 49*(3), 2536–2543.

Watson, G. (2012). The insanity defense. In A. Marmor (Ed.), *Routledge companion to philosophy of law* (pp. 205–221). New York: Routledge.

Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

69

Jayne Lucke, Brad Partridge, Cynthia Forlini, and Eric Racine

Contents

J. Lucke (✉) • B. Partridge
University of Queensland Centre for Clinical Research, The University of Queensland, Brisbane, QLD, Australia
e-mail: j.lucke@uq.edu.au; b.partridge@uq.edu.au

C. Forlini
Institut de recherches cliniques de Montréal (IRCM), Neuroethics Research Unit, Montréal, QC, Canada

Department of Medicine and Department of Social and Preventive Medicine, Université de Montréal, Montréal, QC, Canada
e-mail: Cynthia.Forlini@ircm.qc.ca

E. Racine
Neuroethics Research Unit, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

Department of Medicine and Department of Social and Preventive Medicine, Université de Montréal, Montréal, QC, Canada

Departments of Neurology and Neurosurgery, Experimental Medicine & Biomedical Ethics Unit, McGill University, Montréal, QC, Canada
e-mail: eric.racine@ircm.qc.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_72,

Abstract

This chapter provides an overview of policy and regulatory issues relating to the use of neuropharmaceuticals for cognitive enhancement in normal persons without a cognitive disorder. It draws on experience with a range of policy and regulatory approaches to alcohol, tobacco, pharmaceutical drugs, and illicit drugs and focuses on approaches that target rates of drug use in the population as a whole. The focus on regulatory interventions for the control of neuropharmaceuticals is important because a range of pharmaceutical drugs is often reportedly used and advocated for enhancement purposes. We also examine how more public health interventions such as awareness raising, education, and stigmatization could be used as preventive strategies to reduce the use, and harm associated with the use, of neuropharmaceuticals for cognitive enhancement.

Introduction

The recent increase in neuroscience research on human cognitive functioning has prompted speculation about the potential for developing new interventions, such as neuropharmaceuticals, that will enhance cognitive function in persons whose cognitive performance is normal, that is, unimpaired by disease or injury. This speculation has been encouraged by research reporting improved cognitive performance in "healthy controls" in response to drugs intended for use in the treatment of cognitively impaired patients, e.g., persons who have suffered strokes or head trauma.

This chapter provides an overview of unique policy and regulatory issues raised by the proposed use of pharmaceutical drugs for the purpose of cognitive enhancement in persons without cognitive impairment. Current discussions about the possibility of this type of cognitive enhancement focus on the use of neuropharmaceuticals to improve executive function, memory, attention span, concentration, and alertness. The literature cites some empirical evidence that students, particularly those at elite colleges in the USA, use prescription stimulants such as Ritalin (methylphenidate) or dexamphetamine/mixed amphetamine salts (Adderall) as study aids (DeSantis et al. 2008; Teter et al. 2006). Illicit drugs may also be used for cognitive enhancement purposes, but this chapter will focus on policy and regulatory interventions for the control of neuropharmaceuticals. This is because there is an increasing call for the prescription of neuropharmaceuticals for cognitive enhancement purposes. However, such a policy change could trigger a rapid growth in drug use that may lead to unanticipated harm including addiction. This chapter draws on experience with a range of policy and regulatory approaches to a variety of psychoactive substances that include alcohol, tobacco, pharmaceutical, and illicit drugs. An assessment of different approaches to controlling or regulating

these substances can provide useful insights in designing practical regulatory approaches to the emerging enhancement use of neuropharmaceutical drugs.

Neuropharmaceuticals putatively used for cognitive enhancement can be framed in different ways that reflect similarities between their patterns of use and the use of illicit drugs, prescription drugs, and "lifestyle" drugs. Different fields use different competing frames or discourses to capture the phenomenon. These range from the paradigm of drug misuse and prescription drug abuse (that is often encountered in the public health literature) to the paradigm of "cognitive enhancement" most often encountered in interdisciplinary bioethics discussions (Forlini and Racine 2012; Racine 2008). We also examine the possible roles of public health interventions such as awareness raising, education, and uses of stigma that could be used to prevent use and reduce harm arising from neuroenhancement use of these drugs.

Claims Made for the Enhancement Use of Neuropharmaceuticals

The safety and effectiveness of using neuropharmaceuticals as cognitive enhancers in healthy individuals remains uncertain, but this pattern of use has nonetheless been sympathetically discussed in leading scientific (*Nature*) and medical journals (*BMJ*, *Neurology*). This proposed use of prescription medications raises regulatory and public health policy challenges that need to be carefully considered.

The lack of scientific evidence on the safety and efficacy of neuropharmaceuticals when used for enhancement presents a major challenge to those seeking to make recommendations about policy and regulatory approaches. Nonetheless, professional bodies have attempted to develop policies in the UK (Academy of Medical Sciences 2008; British Medical Association 2007) and Canada (Government of Quebec 2009). The American Academy of Neurology, for example, has published guidance for neurologists on how to respond to patient requests for a prescription to improve their memory, cognitive focus, or attention span, on the assumption that such requests are already common in neurology practice (Larriviere et al. 2009).

Regulatory Options

A recent paper in the Lancet argued that the aim of drug policy should be to "promote the public good by improving individual and public health, neighbourhood safety, and community and family cohesion, and by reducing crime" (Strang et al. 2012, p. 71). Models of drug regulation vary according to the type of drug and the jurisdictions involved. This section examines the strengths and weaknesses of applying these different models of regulation to the use of neuropharmaceuticals for cognitive enhancement. Table 69.1 provides an overview of approaches and their major strengths and weaknesses.

Table 69.1 Summary of regulatory and public health approaches and their major strengths and weaknesses[a]

Approach	Description	Strengths	Weaknesses
Free market	No restrictions on the adult use of neuropharmaceuticals for cognitive enhancement	Promotes personal choice to engage in cognitive enhancement	Potential dangers if safety has not been established
			Resources wasted if efficacy is not established
			Incompatible with current regulation of putative enhancers
Market regulation	Psychoactive substances regulated through various measures including taxes and limits on sales and promotion	Avoid the potential negative consequences of prohibition while controlling use	Commercial interests selling these drugs have an interest in limiting the effectiveness of the system to maximize profits
Licensed users	People demonstrate an understanding of the risks and capacity to use the drugs responsibly before being allowed to use for enhancement	Promotes informed consent	Not currently feasible
			Challenging given varying levels of health literacy in society
Prescription system	Only licensed physicians can prescribe a drug which must be dispensed by a pharmacist	Supervision by health-care professionals	Possibility of diversion
		Ensures access for approved medical purposes	Availability through Internet pharmacies which are difficult to regulate
Prohibition	Unauthorized trade in psychoactive substances, whether illicit or medicinal, is a criminal offence	Clear regulation on nonmedical uses of medications	Creation and expansion of black markets
			Creates challenges because current putative enhancers are used as treatment for legitimate medical conditions
Public health education approaches	Reduce or delay initiation and prevent regular and dependent drug use	Cost-effective	Not very effective for prevention of use of illicit drugs, tobacco, and alcohol
Mass media and information campaigns	Use warning messages about the health impact of drug to influence attitudes and behaviors	Potential for widespread diffusion of information	Potential source of misinformation
Stigmatization and denormalization	Use of stigma to create a negative image for use of neuropharmaceuticals for cognitive enhancement	Not restricted to health agencies	Unintended consequences and collateral stigma of illness
		Can involve social control that is exerted by family and friends	Potential impact on users' self-esteem and societal acceptance of diversity

[a]See text for full details

Free Market Approaches

Some bioethicists have argued for a laissez-faire approach towards neuroenhancement. They have recommended a free market, suggesting that there should be no restrictions on the use of neuropharmaceuticals for cognitive enhancement (Sandberg and Savulescu 2011; Savulescu and Bostrom 2009). Greely and colleagues argued that easier access to cognitive enhancers should be allowed and no legal penalties be imposed on those who wanted to use neuropharmaceuticals without a prescription. Furthermore they suggest that the use of drugs should be viewed "in the same general category as education, good health habits, and information technology – ways that our uniquely innovative species tries to improve itself" (Greely et al. 2008, p. 702).

Bostrom and Sandberg have advocated a similarly liberal view in suggesting that treating nontherapeutic neuropharmaceutical use as misuse is inconsistent with the shifting border between therapy and enhancement. Bostrom and Sandberg (2009) argue that:

> To make the best use of new opportunities, society needs a culture of enhancement, with norms, support structures, and a lay understanding of enhancement that takes it into the mainstream cultural context. Consumers also need better information on risks and benefits of enhancers, which suggests a need for reliable consumer information and for more studies to determine safety and efficacy. (p. 333)

This liberal approach is inconsistent with international drug control treaties. The United Nations Drug Conventions are clear that unauthorized trade in psychoactive substances, whether illicit or medicinal, should be a criminal offence and consequently recreational drug use must be strictly controlled (Hughes and Winstock 2012). Furthermore, free market approaches are inconsistent with other regulatory approaches to the regulation of pharmaceutical drugs in developed countries, as we discuss below. The use of prescription neuropharmaceuticals without a prescription – including those used for cognitive enhancement – is illegal in most developed countries that are signatories to these international treaties.

Market Regulation

If the use of neuropharmaceuticals without a prescription was to become legal, the market could be regulated in different ways. For example, markets could be regulated in the same way as markets for other legal drugs such as tobacco or alcohol. The use of alcohol and tobacco is not prohibited and does not require a prescription, but their use is controlled by methods including the imposition of taxes, limiting availability, age limits on use, and restricting promotion of their use and when and where they may be used. These approaches have had reasonable success in limiting availability and controlling the use of some drugs. For example, raising taxes and limiting the availability of alcohol are moderately effective in

reducing levels of alcohol consumption and rates of alcohol-related problems (Babor et al. 2010; Room and Hall 2012). Regulatory controls on smoking in public places, workplaces, bars and restaurants have contributed to reductions in community smoking prevalence (Room and Hall 2012).

A major weakness with a legal regulatory system is that the companies who market the commodities have a major commercial interest in limiting the effectiveness of the system or in using the system to their advantage, e.g., by limiting competition to current market participants ("regulatory capture"). The continued effectiveness of regulatory systems depends on the vigilance and preparedness of governments to operate these systems in the public interest (Room and Hall 2012).

Licenses for Use

Some bioethicists have suggested a licensing system for the use of neuropharmaceuticals for cognitive enhancement. Bostrom and Sandberg have suggested that "enhancement licenses" be issued to people who demonstrate an understanding of the risks and a capacity to use these drugs responsibility. This would mean that users would be able to give informed consent and that any adverse effects of their use could be properly monitored (Bostrom and Sandberg 2009). Dubljevic proposes an "economic disincentives model" whereby users would be licensed, first paying fees for a course about effects and side effects, proving their knowledge by passing an exam, becoming registered as an enhancement user, and obtaining additional medical insurance (Dubljevic 2012).

This idea is not currently feasible under drug regulatory systems in developed countries. It seems reasonable to expect that such a licensing system would require a minimum demonstration of the safety and efficacy of putative enhancers to meet the requirements of informed consent and other criteria of moral acceptability (Racine 2010). It is not clear who should fund research to demonstrate this or whether consent could be free from coercive pressures to use a cognitive enhancer.

Prescription System

Prescription systems are designed to manage the medical use of drugs that have potentially adverse and beneficial health effects. Many illicit drugs were first used for medical reasons, and some illicit drugs (or drugs with similar effects) are still used as medicines, e.g., opioids and stimulants. Prescription monitoring systems can reduce irregular prescribing and patient utilization while allowing patients to access these drugs for appropriate medical purposes (Strang et al. 2012). Under prescription regimens, only licensed physicians can prescribe a drug which must be dispensed by a pharmacist, often in limited quantities to limit diversion. However, prescription regimens do not eliminate nonmedical use of neuropharmaceuticals because these drugs may be diverted to nonmedical use, e.g., by giving drugs to family or friends or selling the drug on the black market. This can make reduction of supply through traditional law enforcement difficult (Strang et al. 2012).

The prescription system can allow greater control by regulators. For example, in the USA the 2011 Prescription Drug Abuse Prevention Plan was introduced to reduce misuse and diversion of prescription drugs. The plan includes mandatory education for physicians who prescribe and pharmacies that dispense. Such programs are designed to prevent doctor shopping and drug diversion by allowing physicians and pharmacists to access patient records at the time of writing a prescription or dispensing a drug to ensure that the patient is not receiving multiple prescriptions from different doctors (Holmes 2012).

The emergence of Internet-based pharmacy services makes control of prescription drugs less effective and more challenging (Fischer et al. 2010; Strang et al. 2012). Websites may require a prescription but "rogue" websites may provide drugs without one (Nielsen and Barratt 2009). Online pharmacies are difficult to regulate closely because sites may only be available temporarily, they are difficult to trace, and their operation may cross international boundaries and jurisdictions (Orizio et al. 2011). Real-time monitoring may provide much-needed data about the prevalence of acquiring prescription drugs from online pharmacies and even reduce the overuse of prescription medications. Online monitoring of trends in drug use may also provide better information about broader patterns of "off-label" prescription drug use (Nielsen and Barratt 2009).

Regulation of online pharmacies is challenging, but it is even more difficult to regulate Internet sites where people may access information or supplies of pharmaceutical drugs. For example, social networking websites such as MySpace and Facebook have become a common marketplace for the buying and selling of prescription drugs (Stone and Merlo 2011). There are also websites advising users on how to simulate a disorder in order to persuade a doctor to provide a prescription.

It appears that prescription monitoring systems may reduce overall use, but their impact on reducing diversion or nonmedical use of pharmaceuticals is unclear (Fischer et al. 2010). Diversion is not likely to be effectively targeted by law enforcement approaches. It is likely to be more effective to put in place measures to reduce the overall use of the drug while ensuring that when it is used it has been prescribed appropriately (Fischer et al. 2010).

Strang and colleagues note two weaknesses with the regulation of drugs via a prescription system. Firstly, reduced rates of prescription of some drugs can produce an increased use of other prescription drugs with similar effects, as nonmedical users find alternative drugs to use when distribution of their preferred drug has been restricted. Secondly, restrictions on prescribing have the potential to deny medications to patients who require them for treatment (Strang et al. 2012).

Off-Label Use

Off-label use occurs when a prescription is provided by a physician for a reason other than the licensed purpose of the drug but generally for medical indications. The American Academy of Neurology guidelines adopt a laissez-faire approach that permits off-label use of enhancement, but this view does not capture the fact that off-label therapeutic uses of a drug are warranted under the proviso that physicians are using a drug to treat a patient (Larriviere et al. 2009). The Academy

of Neurology guidelines have been criticized for not providing helpful guidance to health professionals that is consistent with commitments to evidence-based medicine and socially responsible medical practice, and it is also at odds with the general understanding and justification of off-label prescription (Racine and Forlini 2010).

Explicit Enhancement Use of Neuropharmaceuticals

It has been argued that the use of neuropharmaceuticals for cognitive enhancement could be readily dealt with by incorporating explicit enhancement use into existing systems of pharmaceutical regulation (Greely et al. 2008). Bostrom and Sandberg argue that the present system for licensing drugs and medical treatments is overly constraining because it does not allow for the development and marketing of drugs that have a solely enhancing function (Bostrom and Sandberg 2009). They suggest that the "medicine-as-treatment-for-disease" framework creates problems for pharmaceutical companies and users who have to doctor-shop to find a physician who is willing to bend the rules by prescribing a drug for enhancement purposes. Some students admit simulating psychiatric symptoms in order to get prescriptions for drugs that they intend to use for enhancement purposes (Carroll et al. 2006).

Dubljevic proposes a system whereby a government agency would offer a licensing system for pharmaceutical companies to develop drugs that would be used for cognitive enhancement, on the condition that these drugs were first used to treat cognitive disorders (Dubljevic 2012). He suggests that pharmaceutical companies would be interested in this opportunity because of the large potential market for cognitive enhancers. However, given that pharmaceutical companies already have large markets for their products, it is not certain that they would be willing to take on the extra expense and medicolegal risk involved in marketing neuropharmaceuticals to healthy people for cognitive enhancement purposes. Previous attempts to market stimulants for cognitive enhancement have been abandoned (Bell et al. 2012). It is unlikely that pharmaceutical companies would be eager to fund studies to demonstrate the safety and efficacy of drugs to enhance cognitive performance in healthy people. The legal liability involved in providing drugs to healthy people would also be a deterrent because there may be less social tolerance for side effects and adverse events of drugs used electively for nontherapeutic reasons.

Prohibition of All Enhancement Use

In contrast to laissez-faire approaches, prohibition is a policy position which bans enhancement use of a drug. A parallel may be drawn between this approach and the anti-doping policy in competitive sports. In this model, athletes submit to a testing regime targeting a range of identified substances, and there are severe penalties for those who are identified as drug users.

An example of the application of this model to cognitive enhancement could be to drug test students for the use of drugs for cognitive enhancement. However, as in the case of illicit drugs, there is no compelling case for the effectiveness of

school-based drug testing in discouraging drug use by students (Roche et al. 2009). The neuropharmaceuticals used for cognitive enhancement purposes may be exactly the same substances used for the treatment of legitimate medical conditions, and it would therefore be very difficult to work such an anti-doping system in practice. Furthermore, a rigorous testing scheme is no guarantee against the use of performance-enhancing substances.

There are a number of other important reasons that make it hard to justify using drug testing to deter use of cognitive enhancers in educational settings. It may lead to the criminalization of such behavior. The resulting stigma and illegality of such use may deter patients from using these drugs for appropriate medical purposes. It may also unwittingly send the message that drug use is common and in fact normalize use. Furthermore, implementing a drug testing system would be costly and divert resources away from other activities such as education and treatment.

Public Health Approaches

Public health approaches to reducing drug-related harm include measures designed to deter individuals from using substances in ways likely to harm themselves or others. These approaches follow from a policy of prohibiting harmful use of a substance while recognizing that legal and regulatory prohibitions will not prevent all use. Measures to prohibit use or regulate markets, as described above, function to protect public health. However, more general prevention approaches also aim to stop people from starting drug use, delay initiation, or dissuade them from becoming regular and dependent drug users.

Examples of such measures include laws that prohibit driving while under the influence of alcohol or drugs, the public use of drugs, or being intoxicated in public places; public education about the risks of drug use; and the stigmatization of certain drug use or patterns of drug use (and thereby the people who use these drugs in these ways). Public health approaches can be implemented in a variety of settings, such as schools, the mass media, community settings such as the workplace, and in primary health care. Environmental interventions aim to limit the availability of dangerous substances, while educational interventions aim to raise community awareness and knowledge of the adverse effects of drug use.

Knowledge and awareness campaigns are generally ineffective in preventing the use of illicit drugs, tobacco, and alcohol (Babor et al. 2010). More effective strategies include psychosocial developmental interventions (e.g., resilience building programs in schools) and information campaigns to correct community misperceptions (e.g., about the prevalence of use of some types of drugs) (Strang et al. 2012). The latter is particularly relevant in the case of neuropharmaceutical use for cognitive enhancement because US college students perceive the prevalence of this type of drug use as much higher than the actual prevalence (Forlini and Racine 2009; McCabe 2008).

Targeted preventive interventions to curtail stimulant abuse among college students have been recommended. These include efforts to educate students about the dangers of illicit stimulant use, using a denormalizing approach, combined with health education to debunk myths and expose the risks involved while encouraging more appropriate study habits (Rosenfield et al. 2011). One limitation of this approach is that it places the onus on university administrations and neglects the role of physicians in prescribing these drugs (Forlini et al. 2013) and the societal pressures that encourage their use.

The mass media are important vehicles of public information (or misinformation) about the prevalence and risks of drug use. Mass media deserve special attention because of their overestimation of the extent of neuropharmaceutical use for cognitive enhancement (Forlini and Racine 2009; Partridge et al. 2011). Media guidelines may raise awareness of this problem. For example, the Australian Press Council has put forward recommendations on how the mass media should avoid certain messages when reporting on drug use (Australian Press Council 2001). Reports should avoid glamorizing drug use or unintentionally providing information on how to obtain a drug or how to use it in risky ways. Unfortunately these guidelines are not always followed when reporting on enhancement use of neuropharmaceuticals (Partridge et al. 2011).

Stigmatization and denormalization of the use of neuropharmaceuticals for cognitive enhancement deserve more attention given their fundamental role in policy and public health approaches to drug use. Current anti-tobacco campaigns which represent an example of the effects of stigmatizing smokers could tempt policy makers to use a similar approach to enhancement use of neuropharmaceuticals. Room and Hall (2012) argue that the main lesson from global experience with prohibitions of psychoactive substances is that such prohibitions are most successful when there is a strong community norm of abstinence. This approach contrasts with the free market arguments for the normalization (or at least the destigmatization) of cognitive enhancement use of neuropharmaceuticals (Bostrom and Sandberg 2009; Greely et al. 2008).

Denormalization and stigmatization are likely to be difficult approaches to implement in the context of cognitive enhancement in the face of markedly different stakeholder attitudes towards cognitive enhancement. For example, health-care providers condemn the nonmedical use of stimulants for academic performance enhancement, while college students may be more tolerant of such use among their peers (Forlini and Racine 2009). It will be difficult to change the very different mindsets of these groups with stigmatization campaigns.

Challenges in Assessing Regulatory Options

Lack of Evidence on Prevalence, Efficacy, and Harms

The lack of good quality evidence about the cognitive enhancement use of neuropharmaceuticals is a major challenge in assessing regulatory options. Early in the bioethics discourse about cognitive enhancement, there were recommendations

made for policy approaches on the basis of assumptions that the substances under discussion were being used widely across the population and they actually enhanced cognitive performance and were safe to use. However, there is an increasing awareness that such assumptions lack an evidence base and, in fact, may be false (Lucke et al. 2011).

Critics have questioned evidence from US studies of student drug misuse that have been used to justify claims that the use of neuropharmaceuticals for cognitive enhancement is widespread. The interpretation and relevance of these findings for cognitive enhancement has been questioned (Lucke et al. 2011). Other results emerging (Franke et al. 2011) show that the prevalence may be much lower than claimed. Media reports perpetuate the impression that prevalence of use is already high and increasing (Partridge et al. 2011). Data are needed from large-scale focused studies of cognitive enhancement to provide data that will inform appropriate policy development.

Recent reviews of the effects of putative neuroenhancing drugs, such as the stimulants, antidepressants, and acetylcholinesterase inhibitors, have not found convincing support for their efficacy (Lynch et al. 2011; Repantis et al. 2008, 2010a, b). In normal healthy people without any impairment, such as sleep deprivation, stimulants have a very modest impact on memory, and gains are more likely among those with lower baseline ability (Smith and Farah 2011).

Neuropharmaceuticals are regulated because of their potential to adversely affect health. Regular users may develop acute tolerance to the subjective effects of stimulants and often respond by increasing their dose, thereby increasing the risks of toxic side effects and of abuse and dependence. In the USA almost 1 in 20 nonmedical users of prescription stimulant medications meets criteria for stimulant dependence or abuse. Medical complications of acute stimulant intoxication include an altered mental state (from euphoria to psychosis), seizures, and cardiac arrhythmia (Kroutil et al. 2006). Debates about the cognitive enhancement use of neuropharmaceuticals have been criticized for ignoring the potential for abuse and addiction (Bell et al. 2012; Swanson et al. 2011). Concerns about minimizing the potential health effects of neuropharmaceuticals should be paramount in discussions about appropriate policy and regulatory frameworks.

Lack of Evidence of Effectiveness of Methods of Regulation

Drug policy is often contentious, operating in "a complex political terrain characterised by intense controversy, mixed opinion and unrelenting media attention" (Fraser and Moore 2011, p. 505). Ideally, evidence about the effectiveness of policy approaches should be underpinned by good scientific evidence about the effects of the substance in question. But as Strang and colleagues note, policy is often formulated to deal with problems of perceived immediate public importance (such as the emergence of a new type of psychoactive drug or a highly publicized drug-related death of a celebrity), rather than on the basis of scientific knowledge about the prevalence of

the drug's use or its likely harms (Strang et al. 2012). Public debate in drug policy is also driven more by values and politics than by scientific evidence. This context means that even with good evidence, emerging knowledge about neuropharmaceuticals may only be tenuously related to public perception, patterns of use, and approaches to policy and regulation. Evidence about the effectiveness of different policy approaches and their cost-effectiveness can nonetheless help both the public and policy makers to select policies that achieve the goals they desire (Strang et al. 2012).

The rationale for public health approaches may be debatable at this point in time. Outram and Racine have noted that there is a lack of empirical data justifying public health interventions. There is a risk that governments may act on the grounds of probable health risk instead of epidemiological data (Outram and Racine 2011a). Outram and Racine have noted the lack of critical reflection on the underlying rationale for reports that cognitive enhancement practices are rampant. They suggest that these reports constitute important actions from a social standpoint:

> . . . as it currently stands, cognitive enhancement is constituted in a way that challenges the creation of coherent and effective policy recommendations. The different approaches taken to the subject of cognitive enhancement appear to reflect this lack of cohesion. However, policy makers should not simply wait and hope that on balance the benefits turn out to be greater than the risks or to wait for definitional consensus. Some components of cognitive enhancement could be reduced down to clearly identified targets to be further examined. Then, if appropriate, policy should be created that is normative and, amongst other criteria, beneficial to the majority of the population. (Outram and Racine 2011b, p. 323)

Future Directions

So what is the most probable policy approach to the use of neuropharmaceuticals for cognitive enhancement? In the absence of a good evidence base on the safety and efficacy of such drug use, prohibition is likely to be the default position. This position is a precautionary one that aims to reduce the risk of adverse health outcomes.

The neuropharmaceuticals that appear to be used for cognitive enhancement are medicines that are already controlled through prescription systems. Patterns of use described in the bioethics literature as cognitive enhancement are achieved through the diversion of neuropharmaceuticals from their medical use, which is defined as a criminal offence in most countries. Examples include the fraudulent presentation of symptoms to a physician for the purpose of obtaining a diagnosis in order to obtain a prescription for a stimulant, or obtaining neuropharmaceuticals without a prescription, either from a friend or through the Internet. A physician who prescribes a neuropharmaceutical off-label for enhancement purposes to someone without a disorder is also acting illegally.

There is no simple answer to the best way forward in formulating policy approaches to the use of neuropharmaceuticals for cognitive neuroenhancement. Every approach has advantages and disadvantages, but international deliberations about policy must start from the basis of existing systems (Racine and Forlini 2009).

The feasibility and effectiveness of approaches may differ across jurisdictions and vary with the characteristics of the specific neuropharmaceutical under consideration and those who use it. Policy approaches to regulating covert drug use are challenging. It is especially difficult to regulate the use of neuropharmaceuticals which may have been obtained through diverse channels – including a prescription for off-label use that may be given in good faith by a physician or illegally accessed via the Internet without a prescription or manufacture and supply via the black market.

It is important to explore all logically possible options. Potentially useful strategies include targeted public health interventions, professional self-regulation to support existing guidelines for prescription, and developing more effective guidelines for medical professionals. Further examination may be warranted of the potential role that denormalization (or stigma) may play in discouraging enhancement use of neuropharmaceuticals. The framing of any such public discourse will be important if we are to avoid making it more difficult for people to access treatment and obtain help for psychological or health concerns arising from such drug use.

Public debates about the possible benefits and harms of using pharmaceuticals for neuroenhancement provide the best chance of developing sensible policies towards such use (Hall 2004). Meaningful deliberation about the ethical acceptability or regulation of neuroenhancement should not be based on speculation and poor-quality estimates of the prevalence, efficacy, and safety of the use of neuropharmaceuticals for cognitive enhancement. Ideally policies should be evidence based and informed by monitoring and evaluation of their impacts. However, in the absence of such evidence, many important ethical challenges can be addressed within the regulation of the use of neuropharmaceuticals for cognitive enhancement through the existing regulatory systems that govern illicit and pharmaceutical drugs.

Cross-References

- ▶ Attention Deficit Hyperactivity Disorder: Improving Performance Through Brain–Computer Interface
- ▶ Brain Research on Morality and Cognition
- ▶ Drug Addiction and Criminal Responsibility
- ▶ Ethical Implications of Brain Stimulation
- ▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
- ▶ Ethics of Pharmacological Mood Enhancement
- ▶ History of Psychopharmacology: From Functional Restitution to Functional Enhancement
- ▶ Neuroenhancement
- ▶ Reflections on Neuroenhancement
- ▶ Research in Neuroenhancement

▶ Sensory Enhancement
▶ Smart Drugs: Ethical Issues
▶ The Morality of Moral Neuroenhancement
▶ What Is Addiction Neuroethics?

References

Academy of Medical Sciences. (2008). *Brain science, addiction and drugs*. London: United Kingdom Academy of Medical Sciences.

Australian Press Council. (2001). Guideline: Drugs and drug addiction, http://www.presscouncil.org.au/document-search/guideline-drugs-and-drug-addiction/?LocatorGroupID=662&LocatorFormID=677&FromSearch=1. Advisory Guidelines. Accessed 5 Nov 2012.

Babor, T., Caulkins, J., Edwards, G., Fischer, B., Foxcroft, D., Humphreys, K., & Strang, J. (2010). *Drug policy and the public good*. Oxford: Oxford University Press.

Bell, S. K., Lucke, J. C., & Hall, W. D. (2012). Lessons for enhancement from the history of cocaine and amphetamine use. *AJOB Neuroscience, 3*(2), 24–29.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311–341.

British Medical Association. (2007). *Boosting your brainpower: Ethical aspects of cognitive enhancements*. London: BMA.

Carroll, B. C., McLaughlin, T. J., & Blake, D. R. (2006). Patterns and knowledge of nonmedical use of stimulants among college students. *Archives of Pediatrics & Adolescent Medicine, 160*(5), 481–485.

DeSantis, A., Webb, E. M., & Noar, S. M. (2008). Illicit use of prescription ADHD medications on a college campus: A multimethodological approach. *Journal of American College Health, 57*(3), 315–324.

Dubljevic, V. (2012). Toward a legitimate public policy on cognition-enhancement drugs. *AJOB Neuroscience, 3*(3), 29–33.

Fischer, B., Bibby, M., & Bouchard, M. (2010). The global diversion of pharmaceutical drugs non-medical use and diversion of psychotropic prescription drugs in North America: A review of sourcing routes and control measures. *Addiction, 105*(12), 2062–2070.

Forlini, C., & Racine, E. (2009). Autonomy and coercion in academic "cognitive enhancement" using methylphenidate: Perspectives of key stakeholders. *Neuroethics, 2*(3), 163–177.

Forlini, C., & Racine, E. (2012). Added stakeholders, added value(s) to the cognitive enhancement debate: Are academic discourse and professional policies sidestepping values of stakeholders? *AJOB Primary Research, 3*(1), 33–47.

Forlini, C., Gauthier, S., & Racine, E. (2013). Should physicians prescribe cognitive enhancers to healthy individuals? *Canadian Medical Association Journal*. Online ahead of print: 10.1503/cmaj.121508.

Franke, A. G., Bonertz, C., Christmann, M., Huss, M., Fellgiebel, A., Hildt, E., & Lieb, K. (2011). Non-medical use of prescription stimulants and illicit use of stimulants for cognitive enhancement in pupils and students in Germany. *Pharmacopsychiatry, 44*(2), 60–66.

Fraser, S., & Moore, D. (2011). Governing through problems: The formulation of policy on amphetamine-type stimulants (ATS) in Australia. *The International Journal on Drug Policy, 22*(6), 498–506.

Government of Quebec. (2009). Psychotropic drugs and expanded uses: An ethical perspective (Position Statement) Quebec: Commission de l'éthique, de la science et de la technologie.

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456*(7223), 702–705.

Hall, W. (2004). Feeling 'better than well:' Can our experiences with psychoactive drugs help us to meet the challenges of novel neuroenhancement methods? *EMBO Reports, 5*(12), 1105–1109.

Holmes, D. (2012). Prescription drug addiction: The treatment challenge. *Lancet, 379*(9810), 17–18.

Hughes, B., & Winstock, A. R. (2012). Controlling new drugs under marketing regulations. *Addiction, 107*(11), 1894–1899.

Kroutil, L. A., Van Brunt, D. L., Herman-Stahl, M. A., Heller, D. C., Bray, R. M., & Penne, M. A. (2006). Nonmedical use of prescription stimulants in the United States. *Drug and Alcohol Dependence, 84*(2), 135–143.

Larriviere, D., Williams, M. A., Rizzo, M., & Bonnie, R. J. (2009). Responding to requests from adult patients for neuroenhancements: Guidance of the ethics, law and humanities committee. *Neurology, 73*(17), 1406–1412.

Lucke, J. C., Bell, S., Partridge, B., & Hall, W. D. (2011). Deflating the neuroenhancement bubble. *AJOB Neuroscience, 2*(4), 38–43.

Lynch, G., Palmer, L. C., & Gall, C. M. (2011). The likelihood of cognitive enhancement. *Pharmacology, Biochemistry, and Behavior, 99*(2), 116–129.

McCabe, S. E. (2008). Misperceptions of non-medical prescription drug use: A web survey of college students. *Addictive Behaviors, 33*(5), 713–724.

Nielsen, S., & Barratt, M. J. (2009). Prescription drug misuse: Is technology friend or foe? *Drug and Alcohol Review, 28*(1), 81–86.

Orizio, G., Merla, A., Schulz, P. J., & Gelatti, U. (2011). Quality of online pharmacies and websites selling prescription drugs: A systematic review. *Journal of Medical Internet Research, 13*(3), e74.

Outram, S. M., & Racine, E. (2011a). Developing public health approaches to cognitive enhancement: An analysis of current reports. *Public Health Ethics, 4*(1), 93–105.

Outram, S. M., & Racine, E. (2011b). Examining reports and policies on cognitive enhancement: Approaches, rationale, and recommendations. *Accountability in Research, 18*(5), 323–341.

Partridge, B. J., Bell, S. K., Lucke, J. C., Yeates, S., & Hall, W. D. (2011). Smart drugs "as common as coffee": Media hype about neuroenhancement. *PLoS One, 6*(11), e28416.

Racine, E. (2008). Interdisciplinary approaches for a pragmatic neuroethics. *The American Journal of Bioethics, 8*(1), 52–53.

Racine, E. (2010). *Pragmatic neuroethics: Improving treatment and understanding of the mind-brain*. Cambridge, MA: The MIT Press.

Racine, E., & Forlini, C. (2009). Expectations regarding cognitive enhancement create substantial challenges. *Journal of Medical Ethics, 35*(8), 469–470.

Racine, E., & Forlini, C. (2010). Responding to requests from adult patients for neuroenhancements: Guidance of the ethics, law and humanities committee. [Letter]. *Neurology, 74*(19), 1555–1556.

Repantis, D., Schlattmann, P., Lainsey, O., & Heuser, I. (2008). Antidepressants for neuroenhancement in healthy individuals: A systematic review. *Poiesis Prax, 6*, 139–174.

Repantis, D., Laisney, O., & Heuser, I. (2010a). Acetylcholinesterase inhibitors and memantine for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 61*(6), 473–481.

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010b). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*(3), 187–206.

Roche, A. M., Bywood, P., Pidd, K., Freeman, T., & Steenson, T. (2009). Drug testing in Australian schools: Policy implications and considerations of punitive, deterrence and/or prevention measures. *The International Journal on Drug Policy, 20*(6), 521–528.

Room, R., & Hall, W. (2012). Population approaches to alcohol, tobacco and drugs: Effectiveness, ethics and interplay with addiction neuroscience. In A. Carter, W. D. Hall, & J. Illes (Eds.), *Addiction neuroethics* (pp. 247–260). London: Academic Press.

Rosenfield, D., Hebert, P. C., Stanbrook, M. B., Flegel, K., & MacDonald, N. E. (2011). Time to address stimulant abuse on our campuses. [Editorial]. *Canadian Medical Association Journal, 183*(12), 1345.

Sandberg, A., & Savulescu, J. (2011). The social and economic impacts of cognitive enhancement. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities*. Oxford, UK: Blackwell.

Savulescu, J., & Bostrom, N. (Eds.). (2009). *Human enhancement*. Oxford: Oxford University Press.

Smith, M. E., & Farah, M. J. (2011). Are prescription stimulants "smart pills"? The epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological Bulletin, 137*(5), 717–741. 717.

Stone, A. M., & Merlo, L. J. (2011). Attitudes of college students toward mental illness stigma and the misuse of psychiatric medications. *The Journal of Clinical Psychiatry, 72*(2), 134–139.

Strang, J., Babor, T., Caulkins, J., Fischer, B., Foxcroft, D., & Humphreys, K. (2012). Drug policy and the public good: Evidence for effective interventions. *The Lancet, 379*(9810), 71–83.

Swanson, J. M., Wigal, T. L., & Volkow, N. D. (2011). Contrast of medical and nonmedical use of stimulant drugs, basis for the distinction, and risk of addiction: Comment on Smith and Farah. [Comment]. *Psychological Bulletin, 137*(5), 742–748.

Teter, C. J., McCabe, S. E., LaGrange, K., Cranford, J. A., & Boyd, C. J. (2006). Illicit use of specific prescription stimulants among college students: Prevalence, motives, and routes of administration. *Pharmacotherapy, 26*(10), 1501–1510.

Section XV

Human Brain Research and Ethics

Human Brain Research and Ethics

70

Jeremy Sugarman

Contents

Abstract

This section on Human Brain Research and Ethics is comprised of four chapters: ► Chaps. 71, "Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges"; ► 72, "Ethics of Sham Surgery in Clinical Trials for Neurologic Disease"; ► 73, "Research in Neuroenhancement"; and ► 74, "Brain Research on Morality and Cognition". The chapters cover a range of important ethical issues encountered in human brain research. This includes discussions regarding how current understandings of research ethics can be applied to translational clinical research in brain research, the ethical challenges faced when considering research related to neuroenhancement, and the relationship of human brain research to our basic understandings of morality and cognition.

Introduction

Human brain research inevitably involves ethical issues. While in large part these ethical issues are similar to those encountered in research in the life sciences in general, certain issues have been especially poignant in regard to this research. Chapters in this section address a set of these issues across a wide spectrum of human brain research. These include the ethical issues related to

J. Sugarman
Johns Hopkins University, Berman Institute of Bioethics, Baltimore, MD, USA
e-mail: jsugarman@jhu.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_73,

translational neuroscience research, sham surgery in brain research, research in neuroenhancement research, and brain research on morality and cognition.

Translational neuroscience research is directed at ameliorating a spectrum of often devastating brain diseases, which may be associated with a change in cognitive capacity as well as troubling physical manifestations. In many cases, current treatment approaches are inadequate and clinical management is based on symptomatic management. These background conditions are complicated by the fact that many promising therapeutic approaches fail to demonstrate benefit when rigorously evaluated, underscoring the need for careful research before and after instituting therapeutic approaches. In their ▶ Chap. 71, "Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges" Jonathan Kimmelman and Spencer Phillips Hey review some of the basic ethical foundations for translational research (human subjects; integrity; and animal welfare) and then describe some of the ethical issues encountered over the continuum of this research. The continuum of translational research involves several stages: preclinical; early phase and hypothesis exploration; late phase; and post-licensure. Potentially problematic aspects of preclinical research include its design and oversight, bias, external validity, lack of aggregated information across preclinical research, and incomplete reporting. Given the importance of these issues to subsequent phases of translational research, Kimmelman and Hey highlight the importance of addressing their practical and ethical aspects. Early phase trials also raise concerns related to their non-therapeutic intent since many potential participants may have exhausted other therapeutic options. Therefore, potential participants' decision-making may be clouded not only by an underlying condition, but also by desperation, accentuating the need to ensure that consent to participate is actually informed and voluntary. In late phase translational research, the scientific and ethical aspects of selecting a comparator arm (such as a placebo or sham) warrant careful consideration. Finally, post-licensure testing can raise concerns about quality as well as appropriate oversight.

Due to concerns about the validity of brain research in humans that may be profoundly influenced by a placebo effect, sham procedures have been employed in some translational brain research, most notably in the evaluation of promising therapies for Parkinson disease. While using sham procedures may help mitigate some scientific concerns, their use also raises ethical concerns. After reviewing the nature of the placebo effect in assessing outcomes in neurologic disease, ▶ Chap. 72, "Ethics of Sham Surgery in Clinical Trials for Neurologic Disease" Sam Horng and Franklin G. Miller describe the ethical issues and practical implications associated with conducting sham surgery. An important ethical consideration in assessing a particular sham procedure relates to the risk of the procedure itself. Whereas some sham procedures do not pose any substantial risk (other than those potentially related to withholding an intervention that may be helpful), others involve certain harms and risk, such as those that require invasive neurosurgical procedures. Another important ethical dimension of sham surgery is the inherent use of deception in the context of clinical research, highlighting the need for transparency and consent regarding the nature of shams. Nevertheless, as in other translational brain research, obtaining consent may be especially complicated due to the manifestations of an

underlying disease. Horng and Miller conclude their chapter by reviewing an approach to assessing the ethical acceptability of research involving sham, drawing attention to areas that necessitate close attention.

While much human brain research is directed primarily at understanding and ameliorating human disease, another interesting body of research is aimed at neuroenhancement. That is, efforts related to improving the ability to perform mental tasks, activities, and functions, such as memory and cognition. ▶ Chap. 73, "Research in Neuroenhancement" Michael L. Kelly and Paul J. Ford describe and analyze the ethical concerns associated with this type of research. Neuroenhancement research can face familiar ethical issues regarding the assessment of risks and benefits as in biomedical research, but it also raises broad ethical concerns related to the possibility of successful interventions exacerbating social inequities and the long-term social implications that might arise from their use. Further, unlike much clinical translational brain research, neuroenhancement research tends to involve participants who are not ill, thereby shaping the ethical analysis somewhat differently. As such, Kelly and Ford suggest that given the limitations of a biomedical research ethics model in analyzing neuroenhancement research, a "consumer-based model" of ethics should be employed. Although a consumer-based model upholds the need to attend to core ethical concepts such as risks and benefits, it gives wide berth to the goals and values of healthy individuals regarding neuroenhancement research that are not constrained by particular notions of disease. While this nascent model is responsive to a range of ethical concerns in neuroenhancement research, it will require further elaboration and refinement as it faces a series of challenges in application, such as setting boundaries on acceptable research efforts and establishing appropriate oversight.

Finally, in ▶ Chap. 74, "Brain Research on Morality and Cognition" Debra J.H. Mathews and Hilary Bok describe yet another dimension of human brain research and ethics that explores the very basic nature of human morality and cognition. Although a range of approaches from the social sciences have been used in this way, Mathews and Bok rely primarily on findings from brain research using fMRI (functional magnetic resonance imaging), examining the potential relationships of these findings to how we make moral decisions. For example, they review a series of studies investigating the role of emotions in responding to moral dilemmas and others assessing the distinction between actions and omissions. In addition, they take notice of a set of studies analyzing differences among psychopaths and non-psychopaths. While the findings from these studies are intriguing, they are necessarily limited for an array of scientific reasons that should be considered carefully before reaching firm conclusions about the nature of human morality. Moreover, findings about how moral decisions are made do not answer the ethical question about what we ought to do. Given such scientific limitations and the recognition of the current ambit of brain research, Mathews and Bok argue that scientific integrity demands appropriate caution when describing the implications of the findings of this research.

In aggregate, the four chapters in this section provide an overview of many of the ethical issues faced in human brain research, providing a taste of the complexity

faced in the field. Further, each of the chapters not only offers important cautions regarding work in the field that merit attention, but they also offer suggestions for addressing related ethical issues as brain research evolves. Future scholarship should be directed at assessing the effectiveness of these suggestions.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges
▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
▶ Ethical Implications of Cell and Gene Therapy
▶ Ethics of Sham Surgery in Clinical Trials for Neurologic Disease
▶ History of Psychopharmacology: From Functional Restitution to Functional Enhancement
▶ Informed Consent and the History of Modern Neurosurgery
▶ Neuroenhancement
▶ Relationship of Benefits to Risks in Psychiatric Research Interventions
▶ Research in Neuroenhancement
▶ Sensory Enhancement

Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges

71

Jonathan Kimmelman and Spencer Phillips Hey

Contents

Abstract

The entire trajectory of neurological research is rife with uncertainty and risk. Despite the abundance of new molecular targets, therapeutic platforms, and increasingly sophisticated theories of pathophysiology, drug development in neurology remains a failure-prone endeavor. This risk, and its reciprocal opportunity, is thick with ethical and epistemic challenges. Given the low likelihood of demonstrating clinical utility and the devastating impact of neurological disease, what neurological interventions should be investigated? What level of evidence provides a warrant for exposing nonhuman animals and human beings to

J. Kimmelman (✉)
Studies for Translation, Ethics and Medicine (STREAM), Biomedical Ethics/Social Studies of Medicine/Department of Human Genetics, McGill University, Montreal, QC, Canada
e-mail: jonathan.kimmelman@mcgill.ca

S.P. Hey
Biomedical Ethics Unit, McGill University, Montreal, QC, Canada
e-mail: heyspencer@gmail.com

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_130,

burdensome and often invasive strategies? This chapter examines some of the ethical and social dimensions of translating new neurological interventions from laboratory concept into various stages of clinical development, focusing on the dominant ethical and epistemic issues as they emerge (and reemerge) throughout the knowledge production process. It concludes with a brief discussion of several relatively uncharted ethical and epistemic questions as they pertain to CNS clinical translation.

Introduction

The entire trajectory of neurological research – from preclinical to clinical testing – is rife with uncertainty and risk. Despite the abundance of new molecular targets, therapeutic platforms, and increasingly sophisticated theories of pathophysiology, drug development in neurology remains a failure-prone endeavor. In more than one analysis, translation of drugs that target central nervous system (CNS) disorders has the second lowest success rate of any medical domain (Kola and Landis 2004; Pangaloss et al. 2007), and neurological drug developers can point to an ever-expanding casualty list of promising CNS drug candidates that did not make the grade when put to rigorous testing (e.g., NXY-059 for stroke, semagacestat for Alzheimer's, GDNF for Parkinson's).

This risk, and its reciprocal opportunity, is thick with ethical and epistemic challenges. Given the low likelihood of demonstrating clinical utility and the devastating impact of neurological disease (neurological diseases account for approximately 11 % of disability-adjusted life years in high-income countries, cf. WHO 2006), what neurological interventions should be investigated? What level of evidence provides a warrant for exposing nonhuman animals and human beings to burdensome and often invasive strategies?

This chapter examines some of the ethical and social dimensions of translating new neurological interventions from laboratory concept into various stages of clinical development, focusing on the dominant ethical and epistemic issues as they emerge (and reemerge) throughout the knowledge production process. It concludes with a brief discussion of several relatively uncharted ethical and epistemic questions as they pertain to CNS clinical translation.

Ethical Foundations of Clinical Research

Before proceeding, it is important to establish some of the foundational principles governing clinical translation – and some of the distinctive ways CNS drug development threatens the satisfaction of each. From where, precisely, do ethical challenges in CNS drug development stem? What, precisely, is at stake in the way we conduct and organized clinical development in neurology?

Human Subjects

By far, the most visible and regulated source of concerns cluster around the personal interests of human research subjects. Clinical development requires experimenting on human beings. These experiments threaten the welfare of subjects in several ways. The primary aim of trials is not to provide direct benefit to patients, but rather to produce generalizable knowledge that can change the beliefs and practices of physicians and public health systems. Many burdens borne by subjects are thus not primarily aimed at advancing their personal interests, but rather interests that are external to the human subject. Trials furthermore place patients in the hands of clinical investigators whose loyalties are divided between the needs of the individual patients and the goals of scientific understanding.

More fundamentally, some argue that human research is “dehumanizing” or fundamentally degrading in that it treats human being as manipulable biological objects (Jonas 1969). The principles governing human research, which are laid out in the Belmont Report (1979), the Declaration of Helsinki (2011), and several other policies, aim at providing a road map to promoting the social good of research while still protecting the rights and interests of human subjects. In particular, three core principles are widely viewed as governing the protection of human subjects: respect for persons, beneficence, and justice.

CNS translation poses distinctive challenges for designing research in accordance with all three principles. However, it presents particular problems for respect for persons. This principle states that, when patients have decisional capacity, informed consent must be obtained prior to enrollment in a study. Yet many CNS disorders implicate the very organ of decisional capacity. Trials of new CNS drugs must often enroll subjects who either lack capacity or whose capacity may be declining over the course of the experiment.

The majority of human protection policies accommodate research on individuals lacking capacity. For example, the US Federal regulations for protecting human research subjects acknowledge that these vulnerable individuals require “additional safeguards” and proxy consent (45 CFR 46). However, the content of these safeguards and the legal authorization of proxies are not spelled out in any detail. To ensure that these subjects are respected, some authors have proposed the pediatric standard, where research burdens must not exceed minimal risk (Karlawish 2003; Weijer and Miller 2004).

Unfortunately, the minimal risk standard imposes a serious restriction on the acceptable research procedures in CNS drug development. The brain is difficult to access, owing to the blood–brain barrier and anatomical structure, and many neurological diseases are associated with placebo effects. As a consequence, delivery of interventions and/or internally valid study designs often requires approaches, like neurosurgery, sham procedures, lumbar punctures, or exposure to burdensome co-interventions like immunosuppression. Many such procedures are much more than minimally burdensome and nontherapeutic – and are hence impermissible for patients lacking decisional capacity under the pediatric standard.

Integrity of the Research Enterprise

While the three core principles of research ethics apply to many of the issues in CNS drug development, they are not sufficient to protect all of the parties and interests implicated in or affected by clinical translation. Study design, implementation, and reporting should also safeguard the ability of research systems to supply credible, valid, reliable, and accessible evidence to healthcare and public health (London et al. 2012).

One threat to this goal is the high failure rate for new CNS drugs entering clinical translation – especially those addressing neurodegenerative diseases. These failures exact a potentially unjust drain on public research resources. Another perennial threat is the emergence of medical strategies from outside the mainstream of medicine. Because of the very morbid course of many neurological diseases, patient groups sometimes clamor for testing of approaches that lack a firm scientific foundation. Examples include venous angioplasty for multiple sclerosis (Zamboni et al. 2012), use of "patterning" for neurologically impaired children (Ziring et al. 1999), various nonvalidated "stem cell" approaches (cf. Braude et al. 2005), or intravenous immunoglobulin for Alzheimer disease (Schott and James 2012). Although patient subjects may perceive these treatments to be in their best interest, thorough testing of such heterodox therapies expends the limited human and material resources and may mean that development of more promising treatments is delayed.

More generally, just and efficient allocation of research resources remains a somewhat underexplored ethical issue: How should researchers select new candidate therapies for human testing? How rigorously should each candidate be evaluated before it is either advanced into later phases of research or abandoned entirely? The lack of proven effective therapies for many CNS disorders, and the considerable expenses associated with their development, amplifies the need for principled answers to these questions (Hey and Kimmleman n.d.).

Animal Welfare

Neurological translation also requires extensive experimentation on animals, which have moral status. Experiments used in neurological translation are harmful – and often highly morbid. Experiments testing drugs aimed at treating pain must induce pain states in animals; drugs aimed at treating traumatic brain injury start by causing brain injury in animals that have brain anatomies and functions that are similar to human beings. Some commentators (Lafollette and Shanks 1996) question whether any animal experimentation that induces animal suffering for the sake of medical research is ethically justifiable. Other commentators take a less abolitionist approach, viewing its permissibility conditional on variables like limiting human suffering and attainable advances in medical knowledge (Ringach 2011).

Neurological drug development presents distinctive challenges here as well. First, some neurological drugs target disorders that, almost by definition, are conditions of suffering. To develop drugs for diseases like depression, anxiety, or neuropathic pain, researchers must induce depression-like states, anxiety, or neuropathic pain in animals.

Second, many neurological disorders have no clear counterparts in animals. Parkinson disease and Alzheimer disease do not naturally occur in animals. As a consequence, animal models of these diseases have limited value in terms of supporting inferences about clinical utility. This causal disanalogy between the animal experiment and the human trial is a threat to construct validity (Henderson et al. 2013) and renders more fragile the proposition that animal suffering in such circumstances is redeemed by proportionate gains in knowledge.

A third key problem is that, more so than any other drug development area, neurological translation uses animals that have more sophisticated capacity for mentation, e.g., nonhuman primates. This is because many CNS disorders implicate processes – cognition, memory, affect, fine motor skills – that only have counterparts in nonhuman primates. After all, the goal of clinical generalization drives researchers to use animals whose brains have capacities that are as similar as possible to those of human beings. Some commentators draw no moral distinction between induction of suffering in different species (Singer 1990). Others, however, regard higher mentation – for example, capacity for self-awareness – as conferring equal moral consideration (DeGrazia 1999; Weatherall 2006). One set of commentators (Sughrue et al. 2009) explicated what it would entail to extend moral consideration to nonhuman primate stroke studies and concluded that such studies are, in principle, ethically justifiable. But regardless of how one resolves this problem, policies generally extend more stringent protections to animals that have more humanlike mental capacities (several jurisdictions ban all invasive chimpanzee studies, cf. Wadman 2011; the US Institute of Medicine recently issued a report spelling out restrictive criteria for such research cf. Committee on the Use of Chimpanzees in Biomedical Behavioral Research 2011).

A last controversy in CNS research is the creation of human-animal chimeras through neural tissue grafting. Various research efforts aimed at understanding brain function have involved grafting of human neural tissue into the brains of animals – including nonhuman primates. Here, the concern is: Do chimeric animals "acquire" human moral status by virtue of the grafts? If so, is there something transgressive or problematic about bestowing moral status in this way? And ought beings that have acquired humanlike capacities be provided humanlike protections?

To some extent, the answers to these questions hinge on scientific issues. In general, emergence of humanlike traits is a greater concern only where transplantations occur early in brain development, involve large amounts of tissue into more humanlike species (apes as opposed to monkeys), and where tissues are transplanted into brain regions involved in higher brain function (Greene et al. 2005). Whether there is something problematic about bestowing humanlike traits will largely hinge on whether this leads to suffering. Some commentators entertain

the prospect that acquisition of higher brain function could, in principle, lead to flourishing (Ibid.), and at least according to the "equal consideration" view (cf. DeGrazia 1999; Weatherall 2006), nonhuman primates that have acquired humanlike brain functions should be accorded protections similar to those of human beings. The problem, as one commentator points out, is that a laboratory setting, which gives rise to this acquired moral status, is precisely the setting where treatment of the chimera is unlikely to accord with moral principles (Streiffer 2005).

Having surveyed foundational principles governing clinical translation, we now turn to a discussion of key issues arising as interventions are advanced from laboratory conceptions into validated clinical products.

Preclinical

Preclinical testing, which consists of toxicology studies and proof of principle ("animal efficacy studies"), is the first step in transforming laboratory discoveries into clinical application. Well-designed, executed, and reported preclinical studies provide the evidentiary grounds that anchor the ethical basis for launching clinical investigations. Toxicology studies can establish reasonably safe starting doses for first-in-human trials, and they can help researchers to anticipate and manage common toxicities. Efficacy studies provide the rationale for clinical development. Because the clinical utility of new interventions is often not tested until phase 2 trials (see below), the evidence in preclinical efficacy studies must support not merely the first-in-human studies, but all human investigations preceding decisive, "confirmatory" trials (i.e., trials that establish the basis for regulatory approval). Preclinical studies further enable researchers to interpret failed attempts at translation. Clinical trials whose outcomes are discordant with preclinical studies are much more amenable to meaningful interpretation when the former have been well designed, executed, and reported (Kimmelman 2009).

To fulfill the aim of protecting human subjects in subsequent trials, redeeming the sacrifice of animals, and marshalling limited research resources towards the advancement of knowledge, preclinical studies should be designed, executed, and reported in ways that support reliable inferences about an intervention's clinical utility. This obvious proposition, however, runs up against various problems with the way preclinical research activities are currently organized.

For instance, preclinical studies are not held to the same standards of design rigor as typical clinical studies. Whereas drug regulators like FDA hold clinical trials to very stringent quality requirements, drug regulators do not, as a rule, carefully scrutinize preclinical efficacy studies when authorizing clinical development (US FDA 2006). Moreover, ethics review committees, too, are rarely qualified to assess the preclinical studies underwriting clinical development programs. Institutional Animal Care and Use Committees, which review studies for animal welfare, are not empowered to provide stringent review of preclinical studies for their clinical generalizability. As a consequence, investigators and sponsors have

a very wide discretion about the design, execution, and interpretation of preclinical efficacy studies.

In part, because so many CNS drugs that show activity in animals fail clinical development, the neurology translation community has been at the vanguard in addressing sources of bias that may lead to spurious preclinical findings (cf. van der Worp et al. 2005, 2010). These sources fall into four broad categories: First, many preclinical studies do not implement measures to address threats to internal validity, such as randomization and concealed treatment allocation. Second, studies often use deficient representations of human disease or treatment – i.e., a lack of construct validity (Henderson et al. n.d.). As described above, many human neurological illnesses – Parkinson disease, Alzheimer disease, Huntington disease – have no natural counterpart in animals. But other limitations with construct validity are less insuperable. For instance, new drugs are rarely tested in animals that suffer from comorbidities common in human neurological illnesses, even though in some instances, such comorbidities appear to modulate the activity of new drugs (Bath et al. 2009). Interventions are often implemented in preclinical studies that are flatly impractical in trial settings. For instance, drugs for treating stroke must be effective when delivered hours after an ischemic event. However, animal studies often deliver drugs on unrealistic timelines, or they do not run sufficiently long to judge the durability of response (van der Worp et al. 2005).

A third common problem concerns the “external validity” of studies – their capacity to demonstrate that the causal properties of new interventions extend to various untested conditions. In recent years, numerous reports have surfaced documenting problems and deficiencies with replication, quasi-replication, and other attempts to reproduce causal relationships using different methods (Mullard 2011; Begley and Ellis 2012). In Alzheimer disease, for example, early animal studies showed promising results for therapies that targetted the protein fragment amyloid-β, which is known to form plaques in the brains of Alzheimer patients. However, doubts were soon cast on the efficacy of this approach after human trials failed to reproduce the success seen in the animal model, and subsequent animal experiments were unable to ever replicate the early positive finding (Ledford 2010).

A fourth common problem in preclinical research is that it is often not well systematized and aggregated. This reflects a cluster of related problems. For one, systematic reviews are uncommon in preclinical research (Sandercock and Roberts 2002), and as a consequence, human clinical trial brochures, which describe the merits of a proposed study to the ethics review board, do not always capture the totality of relevant evidence. Second, laboratory practices in preclinical research vary widely. This is partly a strength of the research enterprise, because it allows researchers to sample a wide array of conditions and circumstances in their experiments. However, it greatly complicates the process of synthesizing preclinical evidence. In recent years, various commentators have proposed standardizing some of the methods used to perform preclinical research in CNS translation. Many of these proposals are aimed at addressing the three validity threats described above. However, some propose going further by standardizing the animal models and study designs used across different translation efforts (Ludolph et al. 2010; Perel et al. 2007).

Closely related to the design and execution issues canvassed above is the problem of reporting. Underreporting – particularly underreporting of unfavorable or negative findings – occurs in three broad ways. First, researchers sometimes censor their data. For example, animals that show anomalous effects in preclinical studies will often be excluded from the analysis. Second, researchers sometimes opt not to publish their results at all. Study after study has used statistical methods to demonstrate that preclinical efficacy studies in neurological disease are often underreported – and this under reporting biases against negative or null studies (Wheble et al. 2008; Sena et al. 2010). Third, often researchers publish studies but provide very little information about how studies were performed. For example, sample sizes or error bars might not be included in figures (van der Worp et al. 2010).

To some degree, such reporting practices are defensible. Experimental findings have asymmetric epistemic properties where methodologies are not yet well established: Positive treatment effects often have a relatively clear explanation and can be regarded as informative, whereas null or negative results can have myriad plausible explanations and are often not informative at all. For instance, negative results can reflect implementation errors in an experiment, or flaws in design or method, or a general failure to tap into the right causal system. Further to this point, the incentive structures that operate in preclinical research disfavor publication of negative findings. Preclinical research is often conducted in the private sector, where trade secrecy or confidentiality agreements discourage transparency. Even where preclinical studies are performed in academic settings, researchers are not rewarded professionally for publishing negative studies.

Various remedies have been proposed, as for example, guidelines for design of CNS preclinical efficacy studies (STAIR 1999; Fisher et al. 2009; Kilkenny et al. 2010). Other proposals aim at rewarding negative publication, including the establishment of journals that are chartered to publish negative or inconclusive findings (e.g., *PLoS One* or *The All Results Journal*), new methods for systematic review that enable statistical "correction" for underreporting, and establishment of registries and data repositories for preclinical studies (Kimmelman and Anderson 2012). Nevertheless, a crucial challenge here is that improving the protection of human subjects and the clinical research enterprise through better preclinical research may require greater commitment of resources to the latter – and still greater burden on animals. How these burdens and costs in human testing ought to be weighed against those in animal research activities is far from resolved.

Early Phase and Hypothesis Exploration

Clinical development begins with first-in-human trials and extends into other kinds of human trials, including dose-ranging studies, proof of principle studies, biomarker investigations, and preliminary tests of efficacy (e.g., so-called "phase 1" and "phase 2" testing). This stage of clinical translation is generally aimed at exploring and constructing discrete, testable hypotheses concerning clinical utility that can be put to rigorous testing in large, late phase trials.

Early phase studies represent the point of maximum uncertainty concerning the properties of a new intervention. Some of these uncertainties are quantifiable – and significant. For example, the overall risk of permanent or serious neurological deficits due to neurosurgical procedures for delivering agents to deep nuclei in Parkinson disease trials is on the order of 0.8 % (Kimmelman et al. 2011). However, many uncertainties are less bounded in CNS drug development. As previously indicated, some CNS functions – e.g., cognition or affect – are difficult, if not impossible, to model in animal systems. Brain derangements are often irreversible and may not be apparent immediately after receiving an intervention (Duggan et al. 2009).

These uncertainties, which present fundamental problems for protecting subjects and the integrity of the research enterprise, confound two closely related ethical questions: First, how much and what kind of evidence is sufficient to launch human investigations? One proposal is the "principle of modest translational distance," according to which, initial tests in human beings should be scaled and staged according to the levels of uncertainty concerning the biological properties of an intervention (Kimmelman 2009). This means that, where there are significant uncertainties about the pharmacokinetics and pharmacodynamics of a new strategy, initial human investigations might use very limited exposures to a new intervention to nail down one set of variables before advancing to research activities testing safety and efficacy in larger populations. Ultimately, the question of when to initiate human trials is entangled with the intractable problem – discussed below – of what counts as an acceptable risk/knowledge-value balance and how this balance can be improved.

Second, how does one justify exposing volunteers – often cognitively compromised ones – to substantial nontherapeutic risks and burdens with little assurance of therapeutic benefit? Risks of clinical research can potentially be justified by appeal to therapeutic value for subjects or the social value of the knowledge that is expected to accrue from the study. As we will see below, in late phase studies, administration of unproven drugs is supported by evidence of clinical utility from earlier phase studies and hence can credibly claim competitiveness with standard of care. However, such therapeutic justification for risk exposure in early phase studies is suspect on the grounds that, at the stage early phase trials are run, researchers rarely know precisely how to combine the drug with various material and cognitive practices in order to unlock its clinical utility (Kimmelman 2012). Further, low base rates of successful translation suggest that the prior odds that any one CNS drug candidate will have clinical value at the outset of early phase testing are exceedingly small (Kimmelman and London 2011). As a consequence, the ethical justification for administering new CNS interventions in early stages of clinical translation should generally be grounded in knowledge value.

This nontherapeutic orientation of intervention administration in early phase studies has implications for numerous aspects of trial design and conduct, beginning with subject selection: Should trials enroll patients early in a disease process or only after their condition has advanced? Some favor the former, because many interventions are designed to intervene in early stages of neurodegeneration and as such, hold little prospect for benefiting patients in later stages (Lowenstein 2008).

However, patients with recent onset disease are often candidates for other care options or may have a relatively high quality of life. Exposure to untested drug candidates has potentially large opportunity costs for them – especially where neurosurgical procedures are involved.

If the prior odds of clinical utility are low at the outset of investigation, an intervention should initially be tested in patients who will experience lower opportunity costs – namely, patients with advanced disease. However, given that disease might have progressed beyond the window where therapeutic strategies are useful, and that comorbidities will often be present with advanced-disease subjects, this can frustrate the studies' capacity to detect signals of clinical utility and thus adequately to address a key research question for early phase studies.

For informed consent, patients with advanced disease may have exhausted standard treatment, and research has consistently demonstrated that many such patients enter early phase studies primarily motivated by the prospect of therapeutic benefit. Informed consent procedures should strive to explain to subjects that their trial enrollment is unlikely to benefit them medically and is, in fact, more likely to entail burdens. Notwithstanding widely expressed views in bioethics to the contrary, there are grounds for believing that patients often understand the nontherapeutic orientation of research in spite of harboring therapeutic motivations (Kim et al. 2009; Sulmasy et al. 2010; Weinfurt et al. 2008).

The nontherapeutic risk justification for intervention administration in early phases also has implications for many trial design features. Consider the question of whether to randomize patients to new drug on a 1:1 or 2:1 basis, favoring the experimental intervention. Many neurological early phase trials use the former (cf. Cudkowicz et al. 2003, 2006). However, 2:1 randomization requires a larger sample size to attain the same statistical power and therefore exposes greater numbers of subjects to a new and untested agent (Pocock 1979). All else being equal, the principle of risk minimization would favor designs that enroll the fewest patients and hence would discourage use of 2:1 randomization schemes.

Finally, there is the problem of how one designs early phase studies that will produce knowledge sufficient to redeem the burdens and risks of nontherapeutic exposures. There are three ways that early phase trials might produce a gain in knowledge: The first is if the trial produces information that leads to successful translation of the intervention. However, as discussed above, the probability of this kind of knowledge accruing is exceedingly low for many neurological disorders.

A second kind of knowledge is gained if the trial rules out an intervention for further testing, such that research efforts can be redirected toward other therapeutic strategies. Given the high failure rate for new CNS drugs entering clinical testing, the probability of this type of knowledge accruing is high (Kola and Landis 2004). However, if the research community is contemplating a large number of other strategies, the *relative value* of this knowledge is low, since all that is learned is that one of a large number of candidate strategies is not useful.

The third kind of knowledge concerns hypothesis formation and evaluation – in this case – pharmacological or pathophysiological knowledge gained in a study. For example, if researchers observe that a promising agent modulates an intended

target but fails to induce a therapeutic response, researchers have gained a pathophysiological insight – namely, that the relationship between a drug target and clinical response is at best an attenuated one. This knowledge can then help researchers troubleshoot their intervention, or it can help guide decision-making about other strategies that modulate the same target. This last type of knowledge is quite valuable; however, designing studies to produce pharmacological or pathophysiological insights often necessitates additional research procedures like lumbar punctures or biopsies. As noted above, these additional research procedures may not accord with the principles of respect for persons lacking decisional capacity.

A last set of ethical issues concerns the reporting of early phase studies. Reporting of results is the first step through which isolated observations are integrated into a larger body of knowledge, hence redeeming both the sacrifice of volunteers as well as the resources expended in clinical investigation. Reporting is a crucial step in the chain of accruing generalizable knowledge in all phases of research. However, in contrast with later phases, investigators conducting many types of "hypothesis-exploring" studies are not currently required by federal regulations to prospectively register their studies in a public database. As a consequence, many early phase and exploratory studies go unreported. Publication bias in CNS drug development has been reported (Liebeskind et al. 2006; Psaty and Kronmal 2008; Sanossian et al. 2006), though little is known about whether publication biases are higher in early phase, exploratory, or biomarker studies (for one study that paints a reassuring picture of publication practice in neurology diagnosis investigations, see Brazzelli et al. 2009).

Late Phase

Late phase clinical testing generally consists of large trials that use symptomatic or clinical endpoints to confirm the clinical utility of new strategies. Such studies tend to be multicentered, enroll hundreds (or sometimes thousands) of patients, run for a longer observation periods (e.g., 6 months or a year), and involve randomization to comparator treatments. Many ethical issues encountered in early phase research also present themselves in late phase trials – e.g., the problem of subject selection or the challenges of obtaining valid informed consent. However, ethical issues in confirmatory trials have somewhat of a different character. Where the fundamental ethical challenge in early phases is uncertainty, the key challenge for late phase studies is risk (i.e., a more quantifiable and bounded judgment about the probability of undesirable outcomes).

We noted, previously, that risks in research can be ethically justified either by appeal to direct benefits for the subject or benefits to society. In general, risks of new interventions in confirmatory trials draw on the former justification. Yet the goal of producing reliable, clinically relevant knowledge is still present.

By far, the most hotly debated issue in neurology late phase studies is the choice of comparators. Judgments about clinical utility are always comparative. A drug might

moderate a disease course, but if its activity is less than what patients might otherwise receive in care (or if the drug has a more noxious side-effect profile than the alternative), it may still fall short of clinical utility. Accordingly, all trials aimed at demonstrating clinical utility necessitate comparison. In neurology confirmatory testing, comparators are almost always included as an arm of a trial. This is because neurological disorders often have waxing and waning disease courses, endpoints involve subjective measures like functional improvement, and some disorders (e.g., Parkinson disease) are associated with large and durable placebo effects. Though the need for internal comparators is widely accepted, their choice can be controversial.

Many late phase neurology trials use placebo comparators. Under the principle of clinical equipoise – which holds that no patient in trials receives less than standard of care (cf. Freedman 1987; Weijer et al. 2000) – placebo comparators are only ethical where there are no validated care options, where withholding standard of care entails only minimal risk, where patients have opted to decline standard treatments in the course of clinical care, or where justifiable and insuperable material constraints severely limit the supply of an intervention. For example, riluzole is the only validated treatment for amyotrophic lateral sclerosis (ALS), but since its clinical utility is marginal, many ALS patients choose not to receive it. Placebo-controlled trials of ALS can be ethical if they recruit such patients. However, many trials involving neurological disorders use placebo comparators, even though doing so entails that some patients will be deprived of effective treatment (Polman et al. 2008). Such trials violate the principle of clinical equipoise, and because they do not test new treatments head-to-head against standard care options, their results do not provide the kind of comparative information healthcare providers need for choosing between the two treatment options.

Another ethical question arises from the use of sham surgical comparators. We previously noted that the brain is difficult to access. Delivery of many CNS interventions therefore involves surgical interventions. For example, trials testing delivery of neurotrophic factors for treatment of Parkinson disease involved implantation of catheters to deep nuclei (Lang et al. 2006). However, surgical interventions introduce a causal confounder into experiments. Neurosurgery, for example, inadvertently lesions tissues or causes inflammatory processes that can interact with disease symptomatology. In addition, symptomatic diseases like depression, pain, and Parkinson disease are associated with large placebo effects (de la Fuente-Fernández et al. 2002). As a consequence, isolating the causal relationship of a new intervention will often necessitate trials where some subjects receive a sham surgery. A discussion of the ethics of sham surgery is deferred to ▶ Chap. 72, “Ethics of Sham Surgery in Clinical Trials for Neurologic Disease”.

Post-licensure

Production of knowledge about an intervention does not end at the point where a product is introduced into the market and taken up into practice. At the point of product licensure, uncertainties still abound.

For example, a prelicensure phase 3 trial might enroll as many as 1,000 or 2,000 patients. However, a two-arm trial would have to enroll at least 7,000 patients to have 50 % statistical power for detecting a 1 in 500 incidence in a drug-related event that otherwise occurs in the general population with a frequency of 1 in 1,000 patients (Eichler et al. 2008). Therefore, pharmacovigilance – monitoring of drug safety post-licensure – is needed to learn more about safety in larger populations.

There may also be uncertainties about a new intervention's utility in populations or settings that are less controlled than those in clinical trials. For instance, are drugs as effective in senior patients? Are they safe in patients with high blood pressure? These patient groups are often excluded from clinical trials, and therefore, questions may remain about the generalizability of trial findings. Finally, there are questions concerning whether new interventions might be useful in related indications – is a Parkinson disease drug useful for essential tremor? Post-licensure research is an essential component of extending causal inferences to unsampled settings.

The major ethical issues arising in post-licensure research are less about risk and uncertainty for volunteers, because most post-licensure studies are performed in ways that cohere with care standards. "Comparative effectiveness" studies, for example, randomize patients to receive one of two approved drugs, where the aim is not to evaluate efficacy but to compare the net therapeutic benefits of each treatment. For these studies, conventional approaches to research ethics – which are preoccupied with protecting the welfare of subjects – may not highlight the most salient moral issues.

Instead, the overarching ethical problems concern the welfare of patients, physicians, and other actors who depend on the knowledge produced in post-licensure investigations (London et al. 2012). On the one hand is the problem of insufficient research oversight. Unlike prelicensure trials, drug regulators have very little leverage over drug companies concerning the design, implementation, or reporting of post-licensure trials. And because post-licensure investigations entail very little risk to volunteers, ethics review bodies often do not vet them carefully for design quality (in the USA, their regulatory mandate is to evaluate knowledge value against risk to subjects, not to evaluate knowledge value against some other social standard, cf. US DHHS 45 CFR 46). As a consequence, post-licensure trials are occasionally performed or reported in ways that are suspect. This dynamic was readily apparent in the case of the epilepsy drug gabapentin, where litigation revealed that reporting of post-licensure trials was selective and skewed in favor of the drug (Vedula et al. 2009, 2012).

On the other hand is the problem of unduly onerous oversight. Increasingly, public funders, insurers, and healthcare systems are supporting post-licensure, comparative effectiveness trials in non-research settings. In neurology, these trials have looked at such questions as whether carotid stenting performs better than endarterectomy for preventing stroke (Brott et al. 2010). These studies are critical for healthcare systems, because practitioners in non-research settings often perform

differently than their researcher counterparts, and case mixes in practice settings are much more medically diverse than those in research. Moreover, such studies are useful for assessing the cost-effectiveness of treatment strategies.

Comparative effectiveness research often involves practically no departure from care, other than randomized allocation and transfer of medical chart information into a database. Under conventional accounts of research ethics, the fact of randomization would necessitate full independent review, as well as a cumbersome informed consent process. The transaction costs and layer of oversight add substantial logistical, financial, and temporal burdens for comparative effectiveness studies. Some commentators are urging a more permissive approach to comparative effectiveness research (cf. Vickrey et al. 2012). How, precisely, these proposals will promote the social good of comparative effectiveness and pharmacovigilance studies without inadvertently inviting more of the kinds of seeding studies described above remains to be determined.

Conclusion and Frontiers

Many of the ethical and social issues we have discussed pertain not only to new neurological candidates but to all domains of drug development. The shifting of uncertainty to risk as research moves from early to late phase, the need for higher standards in preclinical reporting, and the use of placebo comparators in place of proven, effective treatments – these issues are confronted in many domains of drug development. Nevertheless, translating new neurological interventions from the bench to the bedside does present a distinctive cluster of challenges: The very low base rate for the success of a new drug places extra strain on the preclinical and early clinical phases of development to ameliorate as much uncertainty as possible in advance of human testing. Reduced patient capacity introduces additional complications to the issues surrounding informed consent and subject selection. The invasiveness of some research procedures and delivery approaches presses against long-held views that patients – especially patients debilitated by illnesses that compromise decisional capacity – ought not to be exposed to risky manipulations that lack therapeutic warrant.

While the standard ethical frameworks – built upon the principles of respect for persons, beneficence, and justice – may not definitively resolve these issues, they nevertheless provide a foundation for argument and refinement of CNS research policies. However, there are many other issues concerning the integrity of the research enterprise that are not as well worked out. These include questions about priority setting in clinical development, better protecting parties that draw on knowledge produced in translation, and oversight of post-licensure studies. Here, the standard frameworks for research ethics may not highlight the most salient moral concerns. The risks and uncertainty that persist across all phases of CNS drug development and beyond underline the importance and need for extending and elaborating ethical frameworks to address these problems.

Cross-References

- ▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
- ▶ Ethics in Neurosurgery
- ▶ Ethics of Sham Surgery in Clinical Trials for Neurologic Disease
- ▶ Experimentation in Cognitive Neuroscience and Cognitive Neurobiology
- ▶ Human Brain Research and Ethics
- ▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives
- ▶ Relationship of Benefits to Risks in Psychiatric Research Interventions

References

Bath, P. M., Macleod, M. R., & Green, A. R. (2009). Emulating multicentre clinical stroke trials: A new paradigm for studying novel interventions in experimental models of stroke. *International Journal of Stroke, 4*, 471–479.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Belmont Report. (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. hhs.gov/ohrp/humansubjects/guidance/belmont.html. Retrieved 15 Jan 2013.

Braude, P., Minger, S. L., & Warwick, R. M. (2005). Stem cell therapy: Hope or hype? *British Medical Journal, 330*, 1159–1160.

Brazzelli, M., Lewis, S. C., Deeks, J. J., & Sandercock, P. A. (2009). No evidence of bias in the process of publication of diagnostic accuracy studies in stroke submitted as abstracts. *Journal of Clinical Epidemiology, 62*, 425–430.

Brott, T. G., Hobson, R. W., Jr., Howard, G., Roubin, G. S., Clark, W. M., Brooks, W., Mackey, A., Hill, M. D., Leimgruber, P. P., Sheffet, A. J., Howard, V. J., Moore, W. S., Voeks, J. H., Hopkins, L. N., Cutlip, D. E., Cohen, D. J., Popma, J. J., Ferguson, R. D., Cohen, S. N., Blackshear, J. L., Silver, F. L., Mohr, J. P., Lal, B. K., Meschia, J. F., & CREST Investigators. (2010). Stenting versus endarterectomy for treatment of carotid-artery stenosis. *The New England Journal of Medicine, 363*, 11–23.

Committee on the Use of Chimpanzees in Biomedical and Behavioral Research, National Research Council. (2011). In B. M. Altevogt, D. E. Pankevich, M. K. Shelton-Davenport, & J. P. Kahn (Eds.), *Chimpanzees in biomedical research: Assessing the necessity*. Washington, DC: The National Academy Press.

Cudkowicz, M. E., Shefner, J. M., Schoenfeld, D. A., Brown, R. H., Jr., Johnson, H., Qureshi, M., Jacobs, M., Rothstein, J. D., Appel, S. H., Pascuzzi, R. M., Heiman-Patterson, T. D., Donofrio, P. D., David, W. S., Russell, J. A., Tandan, R., Pioro, E. P., Felice, K. J., Rosenfeld, J., Mandler, R. N., Sachs, G. M., Bradley, W. G., Raynor, E. M., Baquis, G. D., Belsh, J. M., Novella, S., Goldstein, J., Hulihan, J., & Northeast ALS Consortium. (2003). A randomized, placebo-controlled trial of topiramate in amyotrophic lateral sclerosis. *Neurology, 61*, 456–464.

Cudkowicz, M. E., Shefner, J. M., Schoenfeld, D. A., Zhang, H., Andreasson, K. I., Rothstein, J. D., & Drachman, D. B. (2006). Trial of celecoxib in amyotrophic lateral sclerosis. *Annals of Neurology, 60*, 22–31.

de la Fuente-Fernández, R., Schulzer, M., & Stoessl, A. J. (2002). The placebo effect in neurological disorders. *Lancet Neurology, 1*, 85–91.

Degrazia, D. (1999). The ethics of animal research: What are the prospects for agreement? *Cambridge Quarterly of Healthcare Ethics, 8*, 23–34.

Duggan, P. S., Siegel, A. W., Blass, D. M., Bok, H., Coyle, J. T., Faden, R., Finkel, J., Gearhart, J. D., Greely, H. T., Hillis, A., Hoke, A., Johnson, R., Johnston, M., Kahn, J., Kerr, D., King, P., Kurtzberg, J., Liao, S. M., McDonald, J. W., McKhann, G., Nelson, K. B., Rao, M., Regenberg, A., Smith, K., Solter, D., Song, H., Sugarman, J., Traystman, R. J., Vescovi, A., Yanofski, J., Young, W., & Mathews, D. J. (2009). Unintended changes in cognition, mood, and behavior arising from cell-based interventions for neurological conditions: Ethical challenges. *The American Journal of Bioethics, 9*, 31–36.

Eichler, H., Pignatti, F., Flamion, B., Leufkens, H., & Breckenridge, A. (2008). Balancing early market access to new drugs with the need for benefit/risk data: A mounting dilemma. *Nature Reviews. Drug Discovery, 7*, 818–826.

Fisher, M., Feuerstein, G., Howells, D. W., Hurn, P. D., Kent, T. A., et al. (2009). Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke, 40*, 2244–2250.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine, 317*, 141–145.

Greene, M., Schill, K., Takahashi, S., Bateman-House, A., Beauchamp, T., Bok, H., Cheney, D., Coyle, J., Deacon, T., Dennett, D., Donovan, P., Flanagan, O., Goldman, S., Greely, H., Martin, L., Miller, E., Mueller, D., Siegel, A., Solter, D., Gearhart, J., McKhann, G., & Faden, R. (2005). Moral issues of human-non-human primate neural grafting. *Science, 309*, 385–386.

Henderson, V., Kimmelman, J., Fergusson, D., Grimshaw, J., & Hackam, D. (2013). Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. *PLoS Medicine, 10*, e1001489.

Hey, S. P., & Kimmelman, J. (n.d.). Ethics, error, and initial trials of efficacy. *Science Translational Medicine*, (accepted).

Jonas, H. (1969). Philosophical reflections on experimenting with human subject. *Daedalus, 98*, 219–247.

Karlawish, J. H. T. (2003). Research involving cognitively impaired adults. *The New England Journal of Medicine, 348*, 1389–1392.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biology, 8*, e1000412.

Kim, S. Y. H., Schrock, L., Wilson, R. M., Frank, S. A., Holloway, R. G., Kieburtz, K., & de Vries, R. G. (2009). An approach to evaluating the therapeutic misconception. *IRB, 31*, 7–14.

Kimmelman, J. (2009). *Gene transfer and ethics of first-in-human experiments: Lost in translation*. New York: Cambridge University Press.

Kimmelman, J. (2012). A theoretical framework for early human studies: Uncertainty, intervention ensembles, and boundaries. *Trials, 13*, 173 [Epub ahead of print].

Kimmelman, J., & Anderson, J. (2012). Should preclinical studies be registered? *Nature Biotechnology, 30*, 488–489.

Kimmelman, J., & London, A. J. (2011). Predicting harms and benefits in translational trials: Ethics, evidence, and uncertainty. *PLoS Medicine, 8*, e1001010.

Kimmelman, J., Duckworth, K., Ramsay, T., Voss, T., Ravina, B., & Emborg, M. E. (2011). Risk of surgical delivery to deep nuclei: A meta-analysis. *Movement Disorders, 26*, 1415–1421.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews. Drug Discovery, 3*, 711–715.

Lafollette, H., & Shanks, N. (1996). *Brute science: Dilemmas of animal experimentation*. London: Routledge.

Lang, A. E., Gill, S., Patel, N. K., Lozano, A., Nutt, J. G., Penn, R., Brooks, D. J., Hotton, G., Moro, E., Heywood, P., Brodsky, M. A., Burchiel, K., Kelly, P., Dalvi, A., Scott, B., Stacy, M., Turner, D., Wooten, V. G., Elias, W. J., Laws, E. R., Dhawan, V., Stoessl, A. J., Matcham, J.,

Coffey, R. J., & Traub, M. (2006). Randomized controlled trial of intraputamenal glial cell line-derived neurotrophic factor infusion in Parkinson disease. *Annals of Neurology, 59*, 711–715.

Ledford, H. (2010). Key Alzheimer's findings questioned. *Nature, 466*, 1031.

Liebeskind, D. S., Kidwell, C. S., Sayre, J. W., & Saver, J. L. (2006). Evidence of publication bias in reporting acute stroke clinical trials. *Neurology, 67*, 973–979.

London, A. J., Kimmelman, J., & Carlisle, B. (2012). Rethinking research ethics: The case of postmarketing trials. *Science, 336*, 544–545.

Lowenstein, P. R. (2008). A call for physiopathological ethics. *Molecular Therapy, 16*, 1771–1772.

Ludolph, A. C., Bendotti, C., Blaugrund, E., Chio, A., Greensmith, L., et al. (2010). Guidelines for preclinical animal research in ALS/MND: A consensus meeting. *Amyotrophic Lateral Sclerosis, 11*, 38–45.

Mullard, A. (2011). Reliability of 'new drug target' claims called into question. *Nature Reviews. Drug Discovery, 10*, 643–644.

Pangaloss, M. N., Schechter, L. E., & Hurko, O. (2007). Drug development for CNS disorders: Strategies for balancing risk and reducing attrition. *Nature Reviews. Drug Discovery, 6*, 521–532.

Perel, P., Roberts, I., Sena, E., Philipa, W., Briscoe, C., Sandercock, P., Macleod, M., Mignini, L. E., Jayaram, P., & Khan, K. S. (2007). Comparison of treatment effects between animal experiments and clinical trials: Systematic review. *British Medical Journal, 334*, 197–200.

Pocock, S. (1979). Allocation of patients to treatment in clinical trials. *Biometrics, 35*, 183–197.

Polman, C. H., Reingold, S. C., Barkhof, F., Calabresi, P. A., Clanet, M., Cohen, J. A., Cutter, G. R., Freedman, M. S., Kappos, L., Lublin, F. D., McFarland, H. F., Metz, L. M., Miller, A. E., Montalban, X., O'Connor, P. W., Panitch, H., Richert, J. R., Petkau, J., Schwid, S. R., Sormani, M. P., Thompson, A. J., Weinshenker, B. G., & Wolinsky, J. S. (2008). Ethics of placebo-controlled clinical trials in multiple sclerosis: A reassessment. *Neurology, 70*, 1134–1140.

Psaty, B. M., & Kronmal, R. A. (2008). Reporting mortality findings in trials of rofecoxib for Alzheimer disease or cognitive impairment: A case study based on documents from rofecoxib litigation. *Journal of the American Medical Association, 299*, 1813–1817.

Ringach, D. L. (2011). The use of nonhuman animals in biomedical research. *The American Journal of the Medical Sciences, 342*, 305–313.

Sandercock, P., & Roberts, I. (2002). Systematic reviews of animal experiments. *The Lancet, 360*, 586.

Sanossian, N., Ohanian, A. G., Saver, J. L., Kim, L. I., & Ovbiagele, B. (2006). Frequency and determinants of nonpublication of research in the stroke literature. *Stroke, 37*, 2588–2592.

Schott, S., & James, S. D. (2012). Case study: IViG keeps alzheimer's at bay for a decade. *ABC News* Web, July 19.

Sena, E. S., van der Worp, H. B., Bath, P. M., Howells, D. W., & Macleod, M. R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biology, 8*, e1000344.

Singer, P. (1990). *Animal liberation* (2nd ed.). New York: New York Review.

Streiffer, R. (2005). At the edge of humanity: Human stem cells, chimeras, and moral status. *Kennedy Institute of Ethics Journal, 15*, 347–370.

Stroke Therapy Academic Industry Recommendations (STAIR). (1999). Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke, 30*, 2752–2758.

Sughrue, M. E., Mocco, J., Mack, W. J., Ducruet, A. F., Komotar, R. J., Fischbach, R. L., Martin, T. E., & Connolly, E. S. (2009). Bioethical considerations in translational research: Primate stroke. *The American Journal of Bioethics, 9*, 3–12.

Sulmasy, D. P., Astrow, A. B., He, M. K., Seils, D. M., Meropol, N. J., Micco, E., & Weinfurt, K. P. (2010). The culture of faith and hope: Patients' justifications for their high estimations of

expected therapeutic benefit when enrolling in early phase oncology trials. *Cancer, 116*, 3702–3711.
US Department of Health and Human Services (DHHS). (2009). Protection of human subjects. 45 CFR 46.
US Food, and Drug Administration (US FDA). (2006). Guidance for Industry. INDS – Approaches to Complying with CGMP During Phase 1(Draft Guidance). US Department of Health and Human Services.
van der Worp, H. B., de Haan, P., Morrema, E., & Kalkman, C. J. (2005). Methodological quality of animal studies on neuroprotection in focal cerebral ischaemia. *Journal of Neurology, 252*, 1108–1114.
van der Worp, H. B., Howells, D. W., Sena, E. S., Porritt, M. J., Rewell, S., O'Collins, V., & Macleod, M. R. (2010). Can animal models of disease reliably inform human studies? *PLoS Medicine, 7*, e1000245.
Vedula, S. S., Bero, L., Scherer, R. W., & Dickersin, K. (2009). Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *The New England Journal of Medicine, 361*, 1963–1971.
Vedula, S. S., Goldman, P. S., Rona, I. J., Greene, T. M., & Dickersin, K. (2012). Implementation of a publication strategy in the context of reporting biases. A case study based on new documents from neurontin litigation. *Trials, 13*, 136.
Vickrey, B. G., Hirtz, D., Waddy, S., Cheng, E. M., & Johnston, S. C. (2012). Comparative effectiveness and implementation research: Directions for neurology. *Annals of Neurology, 71*, 732–742.
Wadman, M. (2011). Animal rights: Chimpanzee research on trial. *Nature, 474*, 268–271.
Weatherall, D. (2006). The use of non-human primates in research: a working group report chaired by Sir David Weatherall. *The Academy of Medical Sciences* [online].
Weijer, C., & Miller, P. B. (2004). When are research risks reasonable in relation to anticipated benefits? *Nature, 10*, 570–573.
Weijer, C., Shapiro, S. H., & Glass, K. C. (2000). Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. *British Medical Journal, 321*, 756–758.
Weinfurt, K. P., Seils, D. M., Tzeng, J. P., Compton, K. L., Sulmasy, D. P., Astrow, A. B., Solarino, N. A., Schulman, K. A., & Meropol, N. J. (2008). Expectations of benefit in early-phase clinical trials: Implications for assessing the adequacy of informed consent. *Medical Decision Making, 28*, 575–581.
Wheble, P. C., Sena, E. S., & Macleod, M. R. (2008). A systematic review and meta-analysis of the efficacy of piracetam and piracetam-like compounds in experimental stroke. *Cerebrovascular Diseases, 25*, 5–11.
World Health Organization (WHO). (2006). *Neurological disorders: Public health challenges*. Geneva: World Health Organization.
World Medical Association (WMA). (2011). WMA Declaration of Helsinki: Ethical principles for medical research involving human subjects. www.wma.net/en/30publications/10policies/b3
Zamboni, P., Galeotti, R., Weinstock-Guttman, B., Kennedy, C., Salvi, F., & Zivadinov, R. (2012). Venous angioplasty in patients with multiple sclerosis: Results of a pilot study. *European Journal of Vascular and Endovascular Surgery, 43*, 116–122.
Ziring, P. R., Brazdziunas, D., Cooley, W. C., Kastner, T. A., Kummer, M. E., Gonzalez de Pijem, L., Quint, R. D., Ruppert, E. S., & Sandler, A. D. (1999). American Academy of Pediatrics. Committee on Children with Disabilities. The treatment of neurologically impaired children using patterning. *Pediatrics, 104*, 1149–1151.

Ethics of Sham Surgery in Clinical Trials for Neurologic Disease

72

Sam Horng and Franklin G. Miller

Contents

Abstract

Controversy over the use of sham surgery has evolved over the last decade. Initially rejected by many bioethicists due to concerns over the level of risk involved and the presumed low prospect of benefit, the use of sham controls has gradually become a standard practice in the mainstream of neurologic clinical trial design. The evolution of these attitudes and the ethical arguments in favor of sham surgery in clinical trials for novel neurologic interventions are reviewed in this chapter, with emphasis on prominent trials in the field of Parkinson disease research. A practical framework for the ethical assessment of sham neurologic

The opinions expressed are those of the authors and do not necessarily reflect the position or policy of the National Institutes of Health, the Public Health Service, or the Department of Health and Human Services.

S. Horng
Department of Neurology, Mount Sinai Medical Center, New York, NY, USA

F.G. Miller (✉)
Department of Bioethics, Clinical Center, National Institutes of Health, Bethesda, MD, USA
e-mail: fmiller@nih.gov

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_75,

procedures is provided, with special concerns pertaining to the cognitive limitations, as well as functional measures used, in populations with neurologic disease.

Introduction

The use of sham procedures, including sham surgery, in clinical trials once raised strong ethical objections, particularly in the setting of Parkinson disease (PD) and the administration of intracranial therapeutics (Macklin 1999; Gillet 2001; Albin 2002; Freed et al. 2001). Many ethicists argued that sham surgery posed an unacceptable level of risk with no counterbalancing prospect of clinical benefit (Weijer 2002). Moreover, sham controls were judged to violate the Hippocratic tenet of “do no harm” (Macklin 1999). These arguments, however, conflated the goals of medical practice with those of clinical research (Horng and Miller 2002). When methodologically necessary, the use of placebo procedures has repeatedly demonstrated the importance of controlling for the expectation effect, a significant therapeutic response that is due to expectation alone (Katsnelson 2011). In the field of PD research, a number of promising novel interventions have been shown through sham controls to contribute no more than placebo to significant clinical improvements (LeWitt et al. 2011; Marks et al. 2010; Gross et al. 2011; Lang et al. 2006). In fact, because of these data, placebo controls have been embraced within the PD research community as an ethically acceptable and methodologically necessary gold standard of clinical trial design (Katsnelson 2011).

The debate of sham surgery was sparked by early experiments using a sham control for the intraputaminal injection of fetal dopaminergic stem cells of PD patients (Freed et al. 2001). Two dueling Sounding Board articles in the *New England Journal of Medicine* argued alternately for the methodological necessity of sham controls and the unacceptability of their risk (Freeman et al. 1999; Macklin 1999).

Throughout the following decade, additional placebo-controlled trials in the PD field, including two gene transfer studies, a cell transplantation trial, and a GDNF infusion trial, demonstrated the importance of a sham control as negative results tempered promising positive data from prior open-label trials (LeWitt et al. 2011; Marks et al. 2010; Gross et al. 2011; Lang et al. 2006). Furthermore, a growing body of placebo-controlled data in a diverse array of other fields has exposed the clinical inefficacy of internal mammary artery ligation for angina, arthroscopic knee surgery for osteoarthritis, and vertebroplasty for pain from vertebral fractures (Beecher 1961; Moseley et al. 2002; Miller et al. 2011).

Placebo controls are important especially in neurology and neurosurgery for several reasons. Outcome measures may involve symptoms assessed with subjective rating scales, including pain, headache, fatigue, mood, cognition, and mobility. A placebo is often necessary to control for the reporting effects of expectation bias. Even seemingly objective measures can be affected by expectation. In PD, for

example, measures of mobility may be subject to bias from both the evaluator and the patient: When patients expect benefit from an intervention, they may be motivated to manifest improved performance. Finally, expectation itself may modulate neurologic function as demonstrated in an emerging body of literature on the placebo effect in a range of diseases, including Alzheimer disease (AD) and PD (Alterman et al. 2011; Benedetti et al. 2006). Belief itself in the therapeutic effect of an intervention has been shown to effect quantitative measures of function, including patterns of brain imaging as well as muscle tone (Benedetti et al. 2005).

Furthermore, neurologic disease can affect cognitive processes which may in turn augment or decrease the placebo effect, requiring a placebo arm to reliably extract the role of expectation from intervention effects. For instance, AD patients with frontal lobe dysfunction have been shown to have diminished expectation-related effects of analgesia in contrast to those with temporal parietal occipital dysfunction alone, who show an intact placebo response (Benedetti et al. 2006). Without a placebo control, one would falsely conclude that analgesia is effective in only those patients with temporal parietal occipital dysfunction, whereas, in truth, the patients with frontal lobe dysfunction have lost the capacity for a positive expectation effect.

Methodological concerns and practical trial design (i.e., shortening trial length and small enrollment) so strongly favor the use of a placebo that it is currently the accepted gold standard in the field of PD, where clinical trials involve the most invasive of placebo control designs: 94 % of investigators approve of a sham control, although only 20 % advocate for higher risk sham procedures involving dural penetration (Kim et al. 2005). Research subjects in these trials also understand the rationale for placebo and demonstrate a deliberation process accounting for risks and benefits when deciding to participate (Kim et al. 2012).

In this chapter, guidelines for assessing the justification of a sham procedure or surgery in a neurological or neurosurgical trial are provided. A survey of research demonstrating the importance of the placebo effect in outcomes of neurologic disease is started with. Next, an ethical framework for the use of a placebo arm is reviewed. Finally, special considerations are examined in the subject populations involved in clinical trials of neurologic disease, particularly neurodegenerative disorders raising questions of competency and decision making.

The Placebo Effect and Neurologic Disease

The justification for the use of invasive sham procedures in evaluating treatments for neurologic disease should be framed within the larger context of clinical research ethics. The goal of clinical research ethics entails the production of useful knowledge that will advance the care of future patients (Emanuel et al. 2000). This goal is held in contradistinction to that of clinical care, which is to optimize the care of an individual patient being treated (Horng and Miller 2002). Concerns over the excessive risk posed by placebo surgery should arguably be considered under a framework specific to the goals of research and not clinical care (Litton and Miller 2005).

In order to ensure that the future implementation of an experimental procedure or neurosurgery would provide benefit to patients without imposing excessive risk or financial cost, a placebo control is often necessary to avoid several sources of false-positive error: (1) bias by unblinded outcome assessors, particularly for outcomes using a subjective rating scale, (2) report bias from the subjects who know what intervention they have received, (3) effects of expectation on outcomes, whether subjective ratings or quantifiable measurements, and (4) confounding effects from any additional factors not attributable to the hypothesized mechanism of the intervention under investigation. In most trials using sham invasive procedures, it will be impossible or unsafe for the physician-investigators performing the procedure to be blinded to the intervention administered; however, those researchers who assess outcomes should be blinded.

Therapeutic outcomes of neurologic disease often include symptoms assessed with subjective rating scales, such as pain, depression, headache, fatigue, quality of life, and mobility/motor function. Expectation can also influence seemingly objective measures such as exercise tolerance or motor function. Finally, quantitative measurements may also be influenced by expectation effects, if a placebo intervention exerts biological effects and modulates the neurologic processes underlying symptoms of the disease.

For example, multiple placebo-controlled trials of acupuncture demonstrate a placebo effect for analgesia, as determined by comparison of the sham acupuncture to a no treatment control, of around 33 % (Horng and Miller 2007). Sham arms for various gene transfer and cell implantation therapies in PD patients revealed improvements in motor function ranging from 13 % to 48 % (Katsnelson 2011). This effect is usually attributed to the placebo effect, although other factors may contribute to observed improvement after receiving a sham intervention. To our knowledge, there have been no sham surgical trials to date with a "no treatment" control group, which would be needed to more precisely assess the improvement that can be attributed to the placebo effect.

Placebo interventions have been shown to modulate a diversity of brain processes, including blood flow to regions of pain processing in healthy subjects given painful stimuli, increased putaminal dopamine release in PD patients, and decreased subthalamic nucleus activity correlating with improvements in motility in PD patients (Wager et al. 2004; Benedetti et al. 2004; De la Fuente-Fernandez et al. 2001).

Expectation may also exert biological effects under "operant conditioning" paradigms, where repeated administrations of an active intervention paired with an inert stimulus induce by association an active response to the inert stimulus over time. Experiments have shown that cyclosporine paired with a flavored drink has been shown to induce a conditioned immunosuppressive response to the flavored drink alone (Goebel et al. 2002). It is conceivable that patients with neurologic disease have similar conditioned therapeutic effects to nontherapeutic aspects of the clinical milieu in research trials.

Finally, nonspecific effects of a therapy such as tissue lesions secondary to electrode placement for deep brain stimulation in epilepsy, PD, or dystonia may

exert therapeutic effects that are optimally controlled under sham conditions (Andrade et al. 2006; De la Fuente-Fernandez et al. 2002). These possible effects would be incidental to the properties of the experimental treatment under investigation and not related to its hypothesized mechanism.

In sum, without a sham control, the expectation effect, as well as investigator and subject biases, may contribute to observed therapeutic effects misattributed to the novel intervention. Especially in placebo-responsive conditions with few treatment alternatives, biased outcome assessment may lead to the erroneous adoption of novel interventions that lack treatment-specific efficacy into standard clinical care, thereby subjecting patients to false claims of benefit, undue risk, and unnecessary financial cost. In fact, invasive procedures are often incorporated into clinical practice without prior rigorous evaluation, owing to weaker US FDA standards (Horng and Miller 2002). There are currently no FDA requirements for surgery unless medical devices are used. Nonetheless, the FDA recently warned patients about the risks associated with a novel extracranial venoplasty and stenting procedure for multiple sclerosis, which has yet to demonstrate proven benefit in randomized, placebo-controlled, double-blinded trials (Kuehn 2012). The controversy over this emerging procedure illustrates the growing concern over rigorous efficacy testing of risky surgeries prior to acceptance into the clinical standard of care.

Ethical Concerns in the Use of Sham Procedures in Neurologic and Neurosurgical Trials

A critical factor in the ethical assessment of sham surgery is the level of risk posed to subjects. Sham pharmaceuticals offer uniformly minimal risk from the procedure itself, although they do pose the risk of withholding treatment when proven effective treatments exist for the condition under investigation. Sham procedures, on the other hand, carry the risk of withholding treatment as well as the additional risk of the procedure itself, which may vary widely along a spectrum. Prior discussions have outlined this spectrum within the field of neurology and neurosurgery and will be reviewed briefly here (Horng and Miller 2007). Trials will require a careful case-by-case evaluation of the risk-benefit ratio of both the active and sham interventional arms (London 2006).

At one end are minimally invasive procedures such as transcranial magnetic stimulation (TMS), radiation, and acupuncture. TMS involves the delivery of magnetic pulses noninvasively to the cortical surface, which decreases cortical excitability. It has been tested in motor function after stroke, as well as depression and epilepsy (Fregni et al. 2006a, b; Sole-Padulles et al. 2006). The sham control consists of noninvasive coil placement and directing of pulses away from the skull. Sham radiation for metastatic brain cancer would similarly involve a noninvasive absence of treatment. Sham acupuncture for any wide variety of painful conditions, including migraine, also presents minimally invasive risk (Linde et al. 2005).

At the other end of the spectrum are invasive procedures in the critical care setting, for example, intracranial vascular bypass surgery or decompressive craniectomy, which may pose risk of great harm and where the risk of no treatment may be high. For intracranial aneurysms at imminent risk for rupture, an artery ligation and laser perforation procedure offers the prospect of bypass with correction of the aneurysm and avoidance of distal vasculature clamping, which poses a risk of ischemia (Grady 2006). Here, the risk of a sham arm would be excessive, as subjects would be subjected to craniotomy and denied intervention for an emergent, life-threatening event. Similarly, decompressive craniotomy for intracerebral hemorrhage is offered to patients at imminent risk of brain herniation. These critically ill patients are at risk of decompensating during the procedure, and, therefore, a sham control would pose excessive risk in the absence of sufficient prospect of benefit. A sham control for a similar procedure in the absence of this emergency might be acceptable, for example, an elective surgical bypass for moyamoya, a form of intracerebral vascular insufficiency which leads to the progressive accumulation of strokes.

Between the two extremes are more invasive interventions, which uniquely allow patients to participate in both sham and intervention arms using a crossover design, such as deep brain stimulator implantation or intracranial drug infusion. Deep brain stimulation (DBS), which modulates the activity of motor circuits in the basal ganglia, is currently approved for use in essential tremor, PD, and dystonia based on the results of placebo-controlled trials (Kupsch et al. 2006). In these trials, DBS devices are implanted into the brain and subjects are tested sequentially under both ON and OFF conditions. The risks of dyskinesia from aberrant stimulation as well as device infection are counterbalanced by the prospect of direct benefit. Subjects in the OFF condition experience delay in the active intervention but are too spared the risks of such treatment. Nonetheless, if the procedures do not work, an invasive procedure has been performed with no benefit and the question remains what to do with the implanted stimulator. This issue was raised in trials for permanent pacemakers for vasovagal syncope, in which the stimulator ON was no better than the stimulator OFF (Connelly et al. 2003).

Intracranial drug infusion, as has been performed in PD subjects, involves the implantation of an intracranial catheter, which carries risk of infection, hemorrhage, and tissue disruption (Nutt et al. 2003; Lang et al. 2006). Sham arms involve saline infusions but allow for crossover to active drug if benefit is subsequently demonstrated. Despite the invasive nature of the control, the opportunity to receive active intervention if beneficial and the initial non-exposure to potential drug toxicity counterbalances the prospect of delayed benefit and risks of catheter implantation.

Somewhat less invasive but not allowing for participation in both experimental and sham arms are intracranial cell implantations and gene transfections (Freed et al. 2001; Olanow et al. 2003; Bjorklund et al. 2003; Prehn et al. 2006; Frank et al. 2005). In the fetal tissue transplantation trials for PD, the sham arm was exposed to a moderately invasive craniotomy with burr holes into the skull not penetrating the dura. Subjects were also exposed to the risks of an immunosuppressive drug regimen and radiation from PET scans to measure study outcomes.

The modified form of placebo craniectomy has since been adopted in multiple subsequent trials with no reported adverse events (Marks et al. 2010; Gross et al. 2011; LeWitt et al. 2011).

It is important to emphasize that prospect of harm is not a priori greater in the placebo arm. As in drugs, the experimental procedure may pose additional risks not present in the sham interventional arm. For instance, intracerebral infusion of various chemotherapeutic or immunomodulatory agents may cause nausea or vomiting. Whole brain radiation poses the risk of radiation necrosis as well as secondary malignancy. Novel interventions in particular may carry additional unanticipated risks that cannot be quantified. For example, gene transfer carries a theoretic risk of inflammatory and/or autoimmune reactions, end-organ damage, malignancy, or death.

In addition, risk is often minimized in the placebo group by designing a sham procedure that ensures double blinding and, in some cases, promotes or allows for a positive expectation in subjects. For example, in the cell implantation trials for PD as discussed above, avoiding the penetration of the dura allowed for the minimization of risk while simultaneously maintaining the double blind and allowing for a positive expectation.

The risk-benefit ratio may in fact be more favorable in the sham procedure arm, if both the sham and the intervention provide significant, comparable placebo responses and there are adverse side effects avoided in the sham arm. A placebo surgery may conceivably pose a clinically significant prospect of benefit without the fully invasive risks of an experimental procedure. A number of PD trials have demonstrated just these results and have sparked debate over the ethics of providing novel interventions which offer significant clinical benefit but have been demonstrated to offer no additional benefit than placebo in head-to-head comparisons. It is critical to consider distinct risk-benefit profiles for both the experimental and sham control arms that would be deemed acceptable in all possible outcomes of the clinical trial results (London 2006).

An additional ethical consideration involves the use of deception and the adequacy of informed consent in sham-controlled placebo procedure trials. If deception is necessary to maintain positive expectations in both the intervention and control arms, it is imperative to explain to subjects the probability that they will receive a placebo control and that the administration of this control may involve deception (Miller et al. 2005; Wendler 1996). Studies have shown that informing prospective subjects about the use of deception does not bias trial outcomes (Weiner and Erker 1986; Martin and Katz 2010).

Furthermore, willing consent to deception does not compromise a respect for subjects or their ability to make autonomous decisions. Some argue that it is further necessary to debrief subjects after the completion of the trial, although feasibility of debriefing in the setting of longitudinal study designs may conceivably allow for this requirement to be waived (Miller and Kaptchuk 2004).

Data from trials involving sham intracranial cell implantations in PD subjects demonstrate that they do in fact understand the rationale, likelihood, and consequences of receiving a placebo surgery (Kim et al. 2012). The majority of subjects

agreed that this design was ethical despite many hoping to receive the active treatment. Thus, it is possible for subjects to understand, consent to, and exercise their personal preferences in a placebo-controlled surgery trial.

Special Consideration in Neurodegenerative Disease: Competency and Decision Making

Subjects with neurodegenerative disease typically experience decline in cognitive function, which may impair their ability to provide informed consent to placebo-controlled research. Depending on the nature of these deficits, subjects may misunderstand the nature or likelihood of receiving sham surgery, may not fully appreciate or rather downplay the risks involved, or may exhibit apathy or lack of motivation to participate. Therefore, it is critical to account for any cognitive differences that might emerge in the study population with a particular assessment of how risk assessment and motivating behaviors may be affected.

An emerging body of research reveals increases in risk-taking behaviors of PD patients on dopaminergic therapy. In card-sorting games, patients with advanced disease exhibit a decreased ability to optimize profits based on odds learning (Maia and Frank 2011). Clinically, patients on high doses of carbidopa-levodopa often acquire a predilection for gambling and high spending. This aspect of the disease raises concern that this population may be at risk for agreeing to placebo-controlled research that is excessively risky or has an unfavorable risk-benefit ratio.

Conversely, patients with advanced AD may exhibit apathy which may prevent them from being willing to participate in research. Additionally, deficits in abstract thinking may prevent them from understanding a placebo-controlled design. In this way, neurodegenerative diseases may predispose patients to unreasonably either reject or solicit research trial participation.

Processes of cognitive decline may also diminish or potentiate the physiologic effects of expectation. As mentioned earlier, Alzheimer disease patients with frontal lobe dysfunction show reduced analgesic response to placebo compared to those with temporal parietal occipital dysfunction (Benedetti et al. 2006). One might speculate that PD patients, or those with dysfunction of dopaminergic circuits in the basal ganglia experience an enhanced expectation effect. Thus, the degree to which the placebo effect is attenuated or enhanced in a particular population as a part of disease pathophysiology might vary the requirement and design of a placebo control.

An extensive body of literature on research in pediatric populations as well as on individuals with developmental delay explores the vulnerability of subjects who cannot provide informed consent (Wendler 1996). There is a general consensus that these subjects require extra protections, particularly in regard to assessment of capacity, calculation of risk-benefit ratio, and acquisition of consent from parties safeguarding the subjects' best interests (London 2006). Standards for assessment may vary or need to be tailored for those subjects who once had capacity and those who never did.

Among various guidelines, the World Medical Association's Helsinki Declaration recommends that (1) the mental condition is a necessary characteristic of the research subjects, (2) the research cannot be equally conducted on those able to provide consent, and (3) that the study addresses a question pertinent to the population subjected to research (58th WMA General Assembly, Seoul, Korea, 2008). The US National Bioethics Advisory Commission additionally recommends that (4) an independent professional assess capacity if the research involves greater than minimal risk and (5) that subjects refusal to participate should always be accepted (Galpern et al. 2012).

An advantage in patients with cognitive decline is that preferences on research participation may be documented early in the course of disease in the form of an advance directive prior to the onset of deficits, although these documents may not be possible or legally binding in some settings (Berghmans 1998; Pierce 2010). An alternative would be to limit the study to those subjects with the condition who are capable of giving informed consent, though excluding those not capable would then limit a more representative sample. One approach might involve initiating the very early efficacy evaluation with those capable of giving informed consent and then expanding the pool in later phase efficacy testing.

Ethical Framework for Placebo-Controlled Neurological Procedure or Neurosurgical Trials

Drawing from existing guidelines on clinical research ethics and outlined in previous reviews (Emanuel et al. 2000; Horng and Miller 2007), we recommend eight benchmarks in the ethical assessment of a placebo-controlled neurologic procedure or neurosurgical trial: (1) The design of the trial addresses an important, clinically relevant question. (2) The placebo arm is necessary to answer this question and there is no scientifically valid alternative design. (3) The risk of the placebo is minimized consistent with the ability to answer the scientific question. (4) The risks are reasonable in light of the prospect of direct benefit and benefit from the knowledge to be gained. (5) Risk-benefit ratio is assessed independently for both the placebo and intervention arms. (6) If there is no prospect of direct benefit to subjects in the placebo arm, the risks are justified by the benefits to future patients of the clinical knowledge to be gained. (7) Deception is authorized by subjects. (8) Decision-making capacity of subjects is adequately assessed and consent provided by an appropriate surrogate if necessary.

Areas where special deliberation is needed in the specifics of a particular trial include: (1) evidence for a placebo response, especially in populations, such as AD patients with frontal lobe dysfunction, who demonstrate decreased expectation bias, and the necessity of a placebo control, (2) minimizing invasiveness of the sham while maintaining expectation effects, (3) level of risk of placebo procedure, (4) opportunity for crossover into the active arm, (5) level of active deception necessary to maintain expectation bias, and (6) nature of decision-making capacity in the setting of cognitive decline.

Conclusion

Over the last 15 years, the use of sham surgery control groups has become accepted as an ethically sound method of evaluating invasive procedures. The use of sham-controlled trials, particularly in the field of Parkinson disease research, has allowed for the objective demonstration that a number of novel interventions provide no greater benefit beyond a sizable placebo effect. In the ethical assessment of these trials, particular attention is made to the risk-benefit ratio of both the placebo and interventional arms, as well as the use of deception and its authorization. Special consideration must also account for aspects of cognitive decline which may impair prospective subjects' understanding of placebo design, ability to make proper risk-benefit assessments, and willingness to engage in research.

Cross-References

▶ Clinical Translation in Central Nervous System Diseases: Ethical and Social Challenges
▶ Deep Brain Stimulation for Parkinson's Disease: Historical and Neuroethical Aspects
▶ Deep Brain Stimulation Research Ethics: The Ethical Need for Standardized Reporting, Adequate Trial Designs, and Study Registrations
▶ Devices, Drugs, and Difference: Deep Brain Stimulation and the Advent of Personalized Medicine
▶ Ethical Implications of Cell and Gene Therapy
▶ Ethics of Functional Neurosurgery
▶ Experimentation in Cognitive Neuroscience and Cognitive Neurobiology
▶ Informed Consent and the History of Modern Neurosurgery
▶ Mental Causation
▶ Neuroimaging Neuroethics: Introduction
▶ Neurosurgery: Past, Present, and Future
▶ Parkinson's Disease and Movement Disorders: Historical and Ethical Perspectives
▶ Risk and Consent in Neuropsychiatric Deep Brain Stimulation: An Exemplary Analysis of Treatment-Resistant Depression, Obsessive-Compulsive Disorder, and Dementia

References

58th WMA General Assembly, Seoul, Korea (2008, October). *WMA declaration of Helsinki – Ethical principles for medical research involving human subjects*. http://www.wma.net/en/30publications/10policies/b3/

Albin, R. L. (2002). Sham surgery controls: Intracerebral grafting of fetal tissue for Parkinson's disease and proposed criteria for use of sham surgery controls. *Journal of Medical Ethics, 28*(5), 322–325.

Alterman, R. L., Tagliati, M., & Olanow, C. W. (2011). Open-label surgical trials for Parkinson disease: Time for reconsideration. *Annals of Neurology, 70*(1), 5–8.

Andrade, D. M., Zumsteg, D., Hamani, C., Hodaie, M., Sarkissian, S., Lozano, A. M., & Wennberg, R. A. (2006). Long-term follow-up of patients with thalamic deep brain stimulation for epilepsy. *Neurology, 66*(10), 1571–1573.

Beecher, H. K. (1961). Surgery as placebo: A qualitative study of bias. *JAMA: The Journal of the American Medical Association, 176*, 1102–1107.

Benedetti, F., Colloca, L., Torre, E., et al. (2004). Placebo-responsive Parkinson patients show decreased activity in single neurons of subthalamic nucleus. *Nature Neuroscience, 7*(6), 587–588.

Benedetti, F., Mayberg, H. S., Wager, T. D., Stohler, C. S., & Zubieta, J. K. (2005). Neurobiological mechanisms of the placebo effect. *The Journal of Neuroscience, 25*(45), 10390–10402.

Benedetti, F., et al. (2006). Loss of expectation-related mechanisms in Alzheimer's disease makes analgesic therapies less effective. *Pain, 121*, 133–144.

Berghmans, R. L. (1998). Advance directives for non-therapeutic dementia research: Some ethical and policy considerations. *Journal of Medical Ethics, 24*(1), 32–37.

Bjorklund, A., Dunnett, S. B., Brundin, P., et al. (2003). Neural transplantation for the treatment of Parkinson's disease. *Lancet Neurology, 2*(7), 437–445.

Connelly, S. J., et al. (2003). Pacemaker therapy for prevention of syncope in patients with recurrent severe vasovagal syncope: Second Vasovagal Pacemaker Study (VPS II): A randomized trial. *JAMA : The Journal of the American Medical Association, 289*(17), 2224–2229.

De la Fuente-Fernandez, R., Ruth, T. J., Sossi, V., Schulzer, M., Calne, D. B., & Stoessl, A. J. (2001). Expectation and dopamine release: Mechanism of the placebo effect in Parkinson's disease. *Science, 293*(5532), 1164–1166.

De la Fuente-Fernandez, R., Schulzer, M., & Stoessl, A. J. (2002). The placebo effect in neurological disorders. *Lancet Neurology, 1*, 85–91.

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? *JAMA : The Journal of the American Medical Association, 283*(20), 2701–2711.

Frank, S., et al. (2005). What is the risk of sham surgery in Parkinson disease clinical trials? A review of published reports. *Neurology, 65*, 1101–1103.

Freed, C. R., Greene, P. E., Breeze, R. E., et al. (2001). Transplantation of embryonic dopamine neurons for severe Parkinson's disease. *The New England Journal of Medicine, 344*(10), 710–719.

Freeman, T. B., Vawter, D. E., Leaverton, P. E., Godbold, J. H., Hauser, R. A., Goetz, C. G., & Olanow, C. W. (1999). Use of placebo surgery in controlled trials of a cellular based therapy for Parkinson's disease. *The New England Journal of Medicine, 341*(13), 988–992.

Fregni, F., Otachi, P. T., Do Valle, A., et al. (2006a). A randomized clinical trial of repetitive transcranial magnetic stimulation in patients with refractory epilepsy. *Annals of Neurology, 60*(4), 447–455.

Fregni, F., Boggio, P. S., Lima, M. C., et al. (2006b). A sham-controlled, phase II trial of transcranial direct current stimulation for the treatment of central pain in traumatic spinal cord injury. *Pain, 122*(1–2), 197–209.

Galpern, W. R., Corrigan-Curay, J., Lang, A. E., Kahn, J., Tagle, D., Barker, R. A., Freeman, T. B., Goetz, C. G., Kieburtz, K., Kim, S. Y., Piantadosi, S., Comstock Rick A., & Federoff, H. J. (2012). Sham neurosurgical procedures in clinical trials for neurodegenerative disease: Scientific and ethical considerations. *Lancet Neurology*, *11*(7), 643–650.

Gillet, G. R. (2001). Unnecessary holes in the head. *IRB: Ethics and Human Research, 23*, 1–6.

Goebel, M. U., Trebst, A. E., Steiner, J., Xie, Y. F., Exton, M. S., Frede, S., Canbay, A. E., Michel, M. C., Heemann, U., & Schedlowski, M. (2002). Behavioral conditioning of immunosuppression is possible in humans. *The FASEB Journal, 16*(14), 1869–1873.

Grady, D. (2006, December 19). With lasers and daring, doctors race to save a young man's brain. *The New York Times*.

Gross, R. E., et al. (2011). Instrastriatal transplantation of microcarrier-bound human retinal pigment epithelial cells versus sham surgery in patients with advanced Parkinson's disease: A double-blind, randomised, controlled trial. *Lancet Neurology, 10*(6), 509–519.
Horng, S., & Miller, F. G. (2002). Is placebo surgery unethical? *The New England Journal of Medicine, 347*, 137–139.
Horng, S. H., & Miller, F. G. (2007). Placebo-controlled procedural trials for neurological conditions. *Neurotherapeutics*, *4*(3), 531–536.
Katsnelson, A. (2011). Experimental therapies for Parkinson's disease: Why fake it? *Nature, 476*(7359), 142–144.
Kim, S. Y., Frank, S., Holloway, R., Zimmerman, C., Wilson, R., & Kieburtz, K. (2005). Science and ethics of sham surgery: A survey of Parkinson disease clinical researchers. *Archives of Neurology, 62*(9), 1357–1360.
Kim, S. Y. H., et al. (2012). Sham surgery controls in Parkinson's disease clinical trials: Views of participants. *Movement Disorders, 27*(11), 1461–1465.
Kuehn, B. M. (2012). FDA warns about the risks of unproven surgical therapy for multiple sclerosis. *JAMA: The Journal of the American Medical Association, 307*(24), 2575–2576.
Kupsch, A., Benecke, R., Muller, J., Trottenberg, T., Schneider, G. H., Poewe, W., Eisner, W., Wolters, A., Muller, J. U., Deuschl, G., Pinsker, M. O., et al. (2006). Pallidal deep-brain stimulation in primary generalized or segmental dystonia. *The New England Journal of Medicine, 355*(19), 1978–1990.
Lang, A. E., Gill, S., Patel, N. K., Lozano, A., et al. (2006). Randomized controlled trial of intraputamenal glial cell line-derived neurotrophic factor infusion in Parkinson disease. *Annals of Neurology, 59*(3), 459–466.
LeWitt, P. A., et al. (2011). AAV2-GAD gene therapy for advanced Parkinson's disease: A double-blind, sham-surgery controlled, randomised trail. *Lancet Neurology, 10*(4), 309–319.
Linde, K., Streng, A., Jurgens, S., et al. (2005). Acupuncture for patients with migraine: A randomized controlled trial. *JAMA : The Journal of the American Medical Association, 293*(17), 2118–2125.
Litton, P., & Miller, F. G. (2005). A normative justification for distinguishing the ethics of clinical research from the ethics of medical care. *The Journal of Law, Medicine & Ethics, 33*(3), 566–574.
London, A. J. (2006). Reasonable risks in clinical research: A critique and a proposal for the integrative approach. *Statistics in Medicine, 25*, 2869–2885.
Macklin, R. (1999). The ethical problems with sham surgery in clinical research. *The New England Journal of Medicine, 341*, 992–996.
Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience, 14*(2), 154–162.
Marks, W. J., Jr., et al. (2010). Gene delivery of AAV2-neurturin for Parkinson's disease: A double-blind, randomised, controlled trial. *Lancet Neurology, 9*(12), 1164–1172.
Martin, A. L., & Katz, J. (2010). Inclusion of authorized deception in the informed consent process does not affect the magnitude of the placebo effect for experimentally induced pain. *Pain, 149*, 208–215.
Miller, F. G., & Kaptchuk, T. J. (2004). Sham procedures and the ethics of clinical trials. *Journal of the Royal Society of Medicine, 97*, 576–578.
Miller, F. G., Wendler, D., & Swartzman, L. C. (2005). Deception in research on the placebo effect. *PLoS Medicine, 2*(9), e262.
Miller, F. G., Kallmes, D. F., & Buchbinder, R. (2011). Vertebroplasty and the placebo response. *Radiology, 259*(3), 621–625.
Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., Hollingsworth, J. C., Ashton, C. M., & Wray, N. P. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *The New England Journal of Medicine, 347*(2), 81–88.
Nutt, J. G., Burchiel, K. J., Comella, C. L., ICV GDNF Study Group, et al. (2003). Randomized, double-blind trial of glial cell line-derived neurotrophic factor (GDNF) in PD. *Neurology, 60*(1), 69–73.

Olanow, C. W., Goetz, C. G., Kordower, J. H., et al. (2003). A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease. *Annals of Neurology, 54*(3), 403–414.

Pierce, R. (2010). A changing landscape for advance directives in dementia research. *Social Science & Medicine, 70*(4), 623–630.

Prehn, A. W., Vawter, D. E., Gervais, K. G., DeVries, R. G., Garrett, J. E., Freeman, T. B., & McIndoo, T. Q. (2006). Studying neurosurgical implants for Parkinson disease: A question of design. *Neurology, 67*(8), 1503–1505.

Sole-Padulles, C., Bartres-Faz, D., Junque, C., et al. (2006). Repetitive transcranial magnetic stimulation effects on brain function and cognition among elders with memory dysfunction. A randomized sham-controlled study. *Cerebral Cortex, 16*(10), 1487–1493.

Wager, T. D., Rilling, J. K., Smith, E. E., et al. (2004). Placebo-induced changes in FMRI in the anticipation and experience of pain. *Science, 303*(5661), 1162–1167.

Weijer, C. (2002). Comment: I need a plaebo like I need a hole in the head. *The Journal of Law, Medicine & Ethics, 30*, 69–72.

Weiner, R. L., & Erker, P. V. (1986). The effects of prebriefing misinformed research participants on their attributions or responsibility. *The Journal of Psychology, 120*, 397–410.

Wendler, D. (1996). Deception in medical and behavioral research: Is it ever acceptable? *The Milbank Quarterly, 74*(1), 87–114.

Research in Neuroenhancement

73

Michael L. Kelly and Paul J. Ford

Contents

Abstract

In this chapter, we argue that framing neuroenhancement research as not necessarily constituting biomedical or behavioral research, as defined by directly investigating disease or ill health, allows for a more robust ethical analysis. The ethics of this kind of research requires a substantially different description from the one offered by a standard biomedical research ethics model. We propose that neuroenhancement research is better described using a non-biomedical, consumer-based model. Such a conceptual shift would clarify ways in which certain types of enhancement research differ from biomedical research, how risks and benefits ought to be weighed, and how such research might be prioritized in the larger health-related research environment. This approach proposes a framework

M.L. Kelly (✉)
Department of Neurosurgery, Cleveland Clinic, Cleveland, OH, USA
e-mail: kellym8@ccf.org

P.J. Ford
Department of Bioethics, NeuroEthics Program, Cleveland Clinic, Cleveland, OH, USA
e-mail: Fordp@ccf.org

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_132,

for ethical analysis that considers possible benefits of enhancement therapy while acknowledging the limitations of the biomedical research ethics paradigm in this setting.

Introduction

Neuroenhancement research is expansive, involving numerous disciplines and fields of inquiry including medical drugs and devices, educational techniques, dietary methods, external computational systems, and even social networks and forms of shared cognition (i.e., the internet), to name only a few. Given this size and diversity, the ethics of neuroenhancement research contains an appropriately wide spectrum of ethical issues, some more controversial than others. While few would argue against the study of enhancements like the neurostimulant effects of coffee, many raise doubts about genetic, surgical, or pharmaceutical enhancement research and technologies, particularly when these efforts do not target disease or promote biomedical health (President's Council on Bioethics 2003). With these concerns in mind, neuroenhancement research is defined in this chapter to include physical, chemical, or biological alterations of individual persons that are not directed to disease or loss of normal function.

Several ethical concerns relate to the pursuit of neuroenhancement research and technologies. First, neuroenhancement research and technology is new without much precedent for the assessment of benefits and risks. Second, the potential efficacy of these technologies could create problems of access and exacerbate social inequalities. Third, regulatory issues abound with regard to the study and approval of such technologies. Fourth, the long-term consequences for human beings and society are unclear (Bostrom and Sandberg 2009).

These issues are fundamentally related to the structure of scientific research under a biomedical model (Bostrom and Sandberg 2009). Biomedical concepts such as health and disease and related ethical concepts such as informed consent frame ethical analysis in research: How risks and benefits are weighed, which endpoints are selected for efficacy in research, what regulations apply, and how long-term outcomes are monitored. How such concepts can be applied to neuroenhancement research projects raises many questions, many of which fall outside the traditional boundaries of a biomedical research ethics model. This chapter will discuss how the limitations of a biomedical research ethics model require a shift toward a consumer-based model, which protects research participants while allowing for the potential benefits of neuroenhancement research.

Research Ethics and the Neuroenhancement Debate

Discussions about neuroenhancement research ethics are often overshadowed by ethical debates about the permissibility of enhancement technologies more

generally (Juengst et al. 2003). These debates focus primarily on the ethical ramifications of a new enhancement technique or drug. For instance, is it ethically permissible to allow college students to use Ritalin for enhanced academic performance? Should the elderly have access to new drugs to reverse some of the memory-adverse effects of aging? Which substances should competitive athletes be allowed to ingest in effort to improve their performance? However, less attention is devoted to how such interventions ought to be studied, what rules should apply, and who should fund such investigations.

Some critics of neuroenhancement technologies might oppose a discussion of research ethics from the very start (Sandel 2004). If using a drug to improve athletic or academic performance is inherently unethical, then researching its side effects or efficacy profile is moot. Research activities would then only promote an already unethical practice. Other critics argue that enhancement research entails unjustifiable risks to the subjects since there are no "health" benefits to such research to balance the risks (Heinz et al. 2012). Such research is also likely to promote undesirable social consequences by exacerbating existing socioeconomic disparities or creating new disparities between those who are willing to risk their health for the sake of enhancement. Finally, at the very least, enhancement research priorities represent a "luxury problem," which cannot make demands on scarce resources needed for funding health-oriented research (Heinz et al. 2012).

These debates are complicated by the fact that the fruits of neuroenhancement research are already pervasive. Computers expand human storage and analytic capacities' test prep courses and innovative learning aides are used to boost academic performance and new memory-enhancing drugs have already been studied in patients with Alzheimer disease (2008). Implantable neural devices such as the "Artificial Hippocampus" have already shown improved memory capacities in mice (Berger et al. 2011). As this volume illustrates, the potential for such research is vast and extends well beyond the traditional indications for disease treatment and prevention.

Even proponents of neuroenhancement research often raise a cautionary flag about the unregulated and unabated pursuit of these technologies (Greely et al. 2008). Careful research is essential for assessing the risks and benefits of interventions while providing for appropriate regulations and safeguards. Even high-risk or harmful enhancement drugs or technologies might benefit from research that makes their use safer. In this way, neuroenhancement research might produce a more traditional benefit for biomedical health, at least broadly construed (Des Jarlais 2000).

The Problematic Biomedical-Enhancement Distinction

Arguments in favor of enhancement research frequently draw comparisons between enhancement interventions and previously established biomedical interventions. For example, Lev et al. argue that vaccines should be considered as a kind of biomedical enhancement. These interventions boost a normal immune system;

enhancing the body's defenses against harmful pathogens. If vaccines "enhance" a normal immune system, then future research investigating corrective surgery for age-related vision loss or memory enhancement for age-related dementia processes would be similarly permissible (Lev et al. 2010). Each of these interventions improves a normal physiologic process for an otherwise healthy individual without generating much ethical controversy.

Of course, vaccines and corrective vision surgery are closely associated with a biomedical concept of disease and health. While normal physiologic processes are enhanced, they are improved in order to combat highly virulent pathogens or debilitating functional deficits. The objective is not to simply enhance a normal human capacity for its own sake (Daniels 2000). Even interventions designed to improve eyesight or memory in the elderly do so in order to prevent recognized diseases such as presbyopia (farsightedness) and/or dementia. While eyesight and memory certainly decline with age, there are established biomedical criteria for deciding when such age-related processes cross a threshold into "disease" and warrant intervention or additional research. Such criteria may be disputed, but the risk-benefit calculus that drives research in these areas relies on recognized biomedical standards to justify the risks of research and intervention (Juengst et al. 2003).

"Dual use" interventions further complicate the biomedical-enhancement distinction. Many potential neuroenhancement interventions were originally studied and approved for biomedical indications (Lev et al. 2010). Pharmaceuticals like modafinil are approved for narcolepsy and sleep-related disorders, but they show promise for cognitive enhancement for soldiers in combat situations and physicians during overnight call (Normann and Berger 2008).

Modafinil has already gained approval for clinical indications, and if the drug can also be used to improve performance, particularly in stressful and socially important situations, why prohibit investigations of these applications? Clinical research has already established the safety and side-effect profiles of such drugs; additional research would only study efficacy among healthy volunteers. Dual use interventions seem to bridge the conceptual gap between biomedical and enhancement-related research and treatment objectives, particularly when a clear social benefit is present.

Problems of Paradigm

The publication of *The Belmont Report* in 1979 set the ethical standard for human subjects research in the United States. The report made explicit the need for proper informed consent, a thorough assessment of risks and benefits, and fair subject selection in all research involving human subjects (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979). These requirements were based on more fundamental ethical principles of respect for persons, beneficence, and justice. The publication of the US Department of Health and Human Services (HHS) policy 45CFR 46 also known as the "Common Rule" codified the Belmont findings into US federal policy, making provisions for

Institutional Review Boards (IRBs) and requirements for consent in biomedical and behavioral research. This policy has even been adopted by several non-health-related federal Departments ranging from the Department of Defense to the Department of Agriculture (U.S. Department of Health and Human Services 2009).

Under this paradigm of biomedical research ethics, several conditions are specific to the ethical practice of biomedical research and presumably neuroenhancement research. Ethical research must (1) have health-related social value, (2) be scientifically valid, (3) use fair subject selection, (4) involve a favorable risk-benefit ratio, (5) be independently reviewed, (6) satisfy informed consent requirements, and (7) respect enrolled participants (Emanuel et al. 2000). Most of these conditions for permissibility are not particularly problematic or controversial for neuroenhancement research projects. There is little dispute about the importance of sound scientific methodology, informed consent, and proper respect of enrolled research participants.

However, requirements for health-related social value and favorable risk-benefit ratio present the most serious challenges to any neuroenhancement research activity (Lev et al. 2010). By definition, neuroenhancement research need not be oriented to disease or ill health. In fact, dual use interventions, such as modafinil, are approved by drug agencies for biomedical indications only; any indication for enhancement is a form of “off-label” use (Bostrom and Sandberg 2009). It is not clear then that the logic of dual use interventions can be applied to drugs or techniques pursued primarily for enhancement purposes; it is even less clear for technologies with no known biomedical indications.

In this way, neuroenhancement technologies cannot claim to possess a medically conceived “health-related social value.” Weighing risks and benefits in such a non-biomedical context requires a fundamentally different calculus. In traditional biomedical research, the additional risks of a research intervention are justified by the harmful effects of disease or ill health. Without direct, primary biomedical applications, neuroenhancement research would seem to lack an appropriate justification for allowing subjects to take on additional risk (Bostrom and Sandberg 2009).

Of course, that is, if one assumes that research should apply strictly to a biomedically conceived notion of health-related social value. The promise of neuroenhancement technology extends well beyond the limits of traditional biomedicine into areas like personality, aging, education, job performance, and athletics, to name a few. A biomedical concept of health relates to such social values only indirectly. The issue then is how to define health and whether a notion of health tied to neuroenhancement requires more than what a biomedical model can offer.

In 1948, the World Health Organization (WHO) defined health as “. . .a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (2003). This expansive definition could easily be construed to include neuroenhancement research and interventions. Cognitive and memory enhancements, for example, show obvious potential benefits for mental, social,

and even physical well-being. Such a definition, however, risks conflating health with other goods such as happiness or success (Kass 1975). A biomedical notion of health relies on concepts such as "sickness" and "abnormal" in order to articulate its goals for research and justify its interventions (Bernard 1927). "Well-being" requires neither a similar notion of normal or abnormal, nor a well-described biomedical research standard by which to measure the success or failure of treatment. Such a non-biomedical description of health fails to explain how research might be pursued safely and effectively.

And yet, many potential neuroenhancement drugs and technologies, such as modafinil, have emerged via biomedical research (Turner et al. 2003). Expanding notions of health and a receptive medical profession have enlarged biomedicine's reach into historically nonmedical areas such as behavioral modification, psychology, and cognitive performance. Research dollars and institutional infrastructure have followed. The result is a modern discussion and approach to health that is dominated by biomedicine to the exclusion of other notions of health and sources of well-being (Kass 1975).

We encounter two major problems for neuroenhancement research under an all-encompassing biomedical paradigm. First, we run the risk of pathologizing the normal. These debates have already been played out very publicly in the most recent revision of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) over controversial diagnoses such as Temper Dysfunctional Disorder with Dysphoria and the Binge Eating Disorder (Cooper 2010). Medicalizing otherwise normal physiologic processes, such as aging or personality, affects more than diagnostic integrity. It disrupts social ties that solidify around care for the elderly (Glannon 2002) and threatens to remove personal responsibility for inappropriate or even illegal behaviors (Kass 2003). Second, we lose the biomedical concept of an "indication" for treatment or research. How much age-related memory loss is abnormal and who decides? It should be made clear that no medical treatment has an indication without a patient's informed consent, but this is not in dispute. Indications in neuroenhancement are set by the individual not biomedicine. There are no broadly accepted concepts of symptoms, signs, diagnosis, disease, and treatment to direct conduct in neuroenhancement research (Juengst et al. 2003).

Toward a Consumer-Based Model of Research Ethics for Neuroenhancement

The limitations of a biomedical research ethics model for neuroenhancement research make clear the need for a new approach. A new model would need to allow for highly subjective notions of risk, benefit, and efficacy within broad constraints of the law. Such a model would also likely require a different notion of informed consent.

In health-related matters, the term "consumer" as opposed to research subject or patient has been controversial. Critics argue that the word "consumer" describes health-related research activities inaccurately as a purely commercial transaction.

The term denies the vulnerability of research participants and the power differential inherent in a physician-patient or researcher-subject relationship (Tallis 1999). The concept has also been controversial and suffers from political associations with unpopular healthcare delivery programs like managed care (Jordan and Court 2010). These critics argue that the resistance to the term "consumer" over time is a strong testament to its danger in health-related encounters.

Proponents of the consumer concept argue that the word "subject" or "patient" is passive and implies a biomedical paternalism of a bygone era. The term "consumer" is even a stronger term than the more recent term "participant" as it reflects the priority of preferences over indications and subjectivity over objectivity. It is also a more complete concept and accounts for the full spectrum of research priorities including non-biomedical areas such as cosmetics, fertility, and lifestyle options (Neuberger 1999).

This debate raises important issues for the ethics of neuroenhancement research. Undeniably, the word consumer has found a place in healthcare today because it serves some function. Enhancement therapies like cosmetic surgery and memory drugs do not conjure up the same notions of vulnerability that justify the use of the word patient, subject, or even participant. Decisions to pursue these interventions are made by healthy individuals with a myriad of goals in mind, health-related or otherwise.

The relationship between a researcher and the research participant or subject is decidedly changed when enhancement is considered. The model for such decision-making is no longer biomedical because the concepts of symptoms, signs, and disease no longer apply. The consumer dictates that indications for the service and the value or "benefit" of the intervention. Research conducted under such a model changes in a number of ways.

First, research priorities are set by consumer demand. Much as any commodity is valued on the basis of consumer demand, the demand for neuroenhancement therapies should drive how research projects are prioritized and funded. Second, benefit must be defined by the consumer and not the researcher or the larger research community. While objective measures exist for assessing performance in, for example, memory or concentration, the value of such measures and our understanding of "efficacy" will be dictated by consumer satisfaction and not by reference to any concept of disease or what is "abnormal." Third, risk must be equitable but not necessarily minimal. Consumer research is subject to lawful restraints, but consumers are allowed more freedom to engage and support research than participants in biomedical research. Hannson and others have argued that the biomedical concept of informed consent is a conceptual failure in the commercial setting due to fundamental differences in the structure of business relationships versus biomedical relationships (Hansson 2006). Fourth, methods in research design must account for consumer feedback. These methods are used in biomedical research, but the difference is the high priority, given such feedback in the enhancement setting. Fifth, researchers have a duty to inform consumers of the latest results and the potential for any harmful effects that may result from the use of neuroenhancement products (National Health and Medical Research Council 2001).

The traditional ethical requirements for biomedical research are helpful to consider but not necessarily required under a consumer model. Requirements such as scientific validity, fair subject selection, independent review, informed consent, and respect for enrolled participants should remain. Under a consumer model, research into neuroenhancement might address a myriad of consumer-driven social values that would not be considered in biomedicine. A favorable risk-benefit ratio takes on new meaning as well. Within constraints of the law, consumers are allowed to take on considerable risk for the sake of many different health and non-health-related benefits. Vulnerable patient populations might still be defined and protected via regulation, but such categories need not incorporate a biomedical definition necessarily.

Potential Hazards and Future Directions

Establishing a consumer model as the basis for neuroenhancement research is not without significant challenges. Ill-defined boundaries between biomedical research and enhancement research will make some neuroenhancement research projects difficult to assess outside of a biomedical regulatory framework. For neuroenhancement research involving human subjects, the role and scope of regulatory oversight committees would be uncertain in many cases. For example, what requirements should exist for informed consent when this concept appears unintelligible in non-biomedical contexts (Hansson 2006)? Finally, who has the authority to determine "appropriate" weighting of risks and benefits when such a calculus is so highly subjective under a consumer-based model?

Professional oversight would also pose a significant hazard for those neuroenhancement technologies requiring professional expertise or approval. Normally, professional conduct falls under several layers of oversight including department heads, professional societies, licensing bodies, institutions, and state and federal governments. Professional activities, research or otherwise, that fall outside of these established authority structures would likely expose research participants to unnecessary and perhaps increased risk.

However, cosmetic surgery might serve as a model for professional neuroenhancement research and practice. Coleman has argued that cosmetic surgery and related research provide a valuable professional service even if indications fall outside of a biomedical model of disease. However, Coleman and others have noted that an attempt by the Dutch government to fund cosmetic surgery as medical care proved unsuccessful due to ambiguities in diagnostic and treatment criteria (Coleman 2006).

Critics have argued that cosmetic surgery, and presumably related research, only reinforces undesirable social norms that ought to be resisted; but these same critics have also fallen short of prohibiting or restricting the activity (Tong and Lindemann 2006). The ethical permissibility of cosmetic surgery and research has remained somewhat unsettled in the literature and without convincing objections, research and practice has continued.

Many concerns about the potential hazards of neuroenhancement research could be answered by turning to commercial examples of regulation. Governmental regulations exist regarding air bags and seat belts in passenger vehicles in the United States based on safety concerns. Bostrom and Sandberg have proposed the idea of "baseline acceptable risk" which would permit the use of neuroenhancement technologies below a publicly established risk level (Bostrom and Sandberg 2009). American federal agencies such as the National Highway Traffic Safety Administration (NHTSA) or the FDA could apply similar regulations to make research and use of neuroenhancement interventions generally safe and effective. Such regulations would help to dispel many concerns raised about coercion, fairness, fraud, and safety in the public use of such technologies (Greely et al. 2008). However, concerns would likely persist regarding the use and distribution of interventions outside of the oversight of traditional biomedical institutional bodies.

Properly regulated, however, there seems to be much gained from a shift to a consumer-based model of clinical research in neuroenhancement, particularly given the recent healthcare economic crisis in the United States. The prioritization of such enhancement research would be driven by consumer demand and corporations, rather than government sources, particularly given the subjective quality of indications for enhancement therapy. This shift in funding sources could avoid debates about utilization of scarce resources for nonmedical ends. If such research yielded medically valuable results, these findings could be used as the basis for initiating a formal medical investigation. This would avoid the costly process of biomedical clinical research approval, and free up regulatory bodies like the IRB from overseeing enhancement research.

It could be argued that much of the high cost associated with medical care in the United States would be unaffected by excluding enhancement-related research. Such research is relatively infrequent, typically not federally funded, and unrelated to the main drivers of high cost in the healthcare system in the United States today. However, making a distinction between biomedical and consumer models may also be useful for clinical practice and healthcare decision-making. As healthcare costs continue to rise in the United States, perhaps, we are asking too much of the disciplines of biomedical research and clinical medicine. Acknowledging the limitations of the biomedical model for improving health allows us to use biomedical research more effectively and places neuroenhancement research in a more coherent conceptual ethical framework.

Conclusion

In this chapter, we argue that an ethics of neuroenhancement research, as defined, requires us to acknowledge the limitations of the biomedical model for research involving human subjects. Much of this discussion turns on the definition of health and what biomedicine can do for such a concept. The goals of neuroenhancement therapy lie outside of a purely biomedical notion of health, although there are some intersections and bidirectional influences. Enhancements might improve memory

or cognition, but they do so absent a biomedical risk-benefit standard or criteria for efficacy. A shift to a consumer-based model of neuroenhancement research ethics would provide a useful framework for evaluation of new interventions. This kind of model acknowledges a more pluralistic notion of health, individual liberty, and human thriving. It also establishes an ethical boundary, subject to modifications, between biomedical and enhancement-related research priorities, which preserves healthcare resources for more strictly medical objectives.

Cross-References

► Brain Research on Morality and Cognition
► Ethical Implications of Brain Stimulation
► Ethics of Brain–Computer Interfaces for Enhancement Purposes
► Gene Therapy and the Brain
► Smart Drugs: Ethical Issues
► The Morality of Moral Neuroenhancement

References

Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

(2008). Retrieved from http://www.clinicaltrials.gov/ct2/show/record/NCT00257673

Berger, T. W., Hampson, R. E., Song, D., Goonawardena, A., Marmarelis, V. Z., & Deadwyler, S. A. (2011). A cortical neural prosthesis for restoring and enhancing memory. *Journal of Neural Engineering*, *8(4)*, 046017-2560/8/4/046017. Epub 2011 Jun 15. doi:10.1088/1741-2560/8/4/046017; 10.1088/1741-2560/8/4/046017.

Bernard, C. (1927). An introduction to the study of experimental medicine. Chapter 2: The a priori idea and doubt in experimental reasoning. In Greene, H. C. (Ed.), (2nd ed., pp. 28–30–76). New York: Dover.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311–341. doi:10.1007/s11948-009-9142-5.

Coleman, S. (2006). A defense of cosmetic surgery. In D. Benatar (Ed.), *Cutting to the core: Exploring the ethics of contested surgeries* (1st ed., pp. 171–182). Lanham: Rowman & Littlefield.

Cooper, A. (2010). *Pacific Standard*. Santa Barbara, CA: Pacific Standard.

Daniels, N. (2000). Normal functioning and the treatment-enhancement distinction. *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees, 9*(3), 309–322.

Des Jarlais, D. C. (2000). Research, politics, and needle exchange. *American Journal of Public Health, 90*(9), 1392–1394.

Emanuel, E. J., Wendler, D., & Grady, C. (2000). WHat makes clinical research ethical? *JAMA: The Journal of the American Medical Association, 283*(20), 2701–2711.

Glannon, W. (2002). Identity, prudential concern, and extended lives. *Bioethics, 16*(3), 266–283.

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature*. doi:10.1038/456702a.

Hansson, S. (2006). Informed consent out of context. *Journal of Business Ethics, 63*, 149–154.

Heinz, A., Kipke, R., Heimann, H., & Wiesing, U. (2012). Cognitive neuroenhancement: False assumptions in the ethical debate. *Journal of Medical Ethics*, *38(6)*, 372–375. doi:10.1136/medethics-2011-100041; 10.1136/medethics-2011-100041.

Jordan, Z., & Court, A. (2010). Reconstructing consumer participation in evidence-based health care: A polemic. *International Journal of Consumer Studies, 34*(5), 558–561. doi:10.1111/j.1470-6431.2010.00906.x.

Juengst, E. T., Binstock, R. H., Mehlman, M., Post, S. G., & Whitehouse, P. (2003). Biogerontology, "anti-aging medicine," and the challenges of human enhancement. *The Hastings Center Report, 33*(4), 21–30.

Kass, L. R. (1975). Regarding the end of medicine and the pursuit of health. *The Public Interest*, *40(40)*, 12–29.

Kass, L. R. (2003). Ageless bodies, happy souls: Biotechnology and the pursuit of perfection. *New Atlantis (Washington, D.C.), 1(1)*, 9–28.

Lev, O., Miller, F. G., & Emanuel, E. J. (2010). The ethics of research on enhancement interventions. *Kennedy Institute of Ethics Journal, 20*(2), 101–113.

Murray, S. (2008). *Safety and efficacy of MEM 1003 versus placebo in patients with mild to moderate Alzheimer's disease*. National Institutes of Health. http://www.clinicaltrials.gov/ct2/show/record/NCT00257673 ed.

National Health and Medical Research Council. (2001). Statement on consumer and community participation in health and medical research. Australia: Commonwealth of Australia.

Neuberger, J. (1999). Do we need a new word for patients? Lets do away with "patients". *BMJ (Clinical Research Ed.), 318*(7200), 1756–1757.

Normann, C., & Berger, M. (2008). Neuroenhancement: Status quo and perspectives. *European Archives of Psychiatry and Clinical Neuroscience, 258 (Suppl 5)*, 110–114. doi:10.1007/s00406-008-5022-2; 10.1007/s00406-008-5022-2.

President's Council on Bioethics. (2003). *Beyond therapy: Biotechnology and the pursuit of happiness*. http://www.bioethics.gov/reports/beyondtherapy/preface.html. Accessed 4 Mar 2013.

Sandel, M. (2004). What's wrong with designer children, bionic athletes, and genetic engineering. *The Atlantic Monthly, 293*, 51–62.

Tallis, R. (1999). Do we need a new word for patients? commentary: Leave well alone. *BMJ (Clinical Research Ed.), 318*(7200), 1757–1758.

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report*. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html). Accessed 1 May 2013.

Tong, H., & Lindemann, H. (2006). Beauty under the knife: A feminist appraisal of cosmetic surgery. In D. Benatar (Ed.), *Cutting to the core: Exploring the ethics of contested surgeries* (1st ed., pp. 183–193). Lanham: Rowman and Littlefield.

Turner, D. C., Robbins, T. W., Clark, L., Aron, A. R., Dowson, J., & Sahakian, B. J. (2003). Cognitive enhancing effects of modafinil in healthy volunteers. *Psychopharmacology, 165*(3), 260–269. doi:10.1007/s00213-002-1250-8.

U.S Department of Health and Human Services. (2009). "Code of federal regulations – title 45 public welfare CFR 46".

Brain Research on Morality and Cognition

74

Debra J. H. Mathews and Hilary Bok

Contents

Abstract

The tools of neuroscience have increasingly been employed to address questions long considered the realm of philosophers, including questions of moral sentiment, choice, cognition, and action. This research seeks to identify the structures within the brain that function in and are the basis of human morality, to improve our understanding, for example, of the role of emotion in moral

D.J.H. Mathews (✉)
Johns Hopkins Berman Institute of Bioethics, Baltimore, MD, USA
e-mail: dmathews@jhu.edu

H. Bok
Johns Hopkins University, Baltimore, MD, USA
e-mail: hbok@jhu.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_133,

decision-making, or the moral failings of psychopaths. If successful, such research might not only allow philosophers to refine their theories based on scientific facts but also might inform social policy.

While research on morality comes from a range of disciplines, including psychology, economics, and the social sciences, this chapter will focus on the subset of this research that uses the tools of neuroscience (primarily fMRI) and is explicit about testing hypotheses related to moral sentiment, choice, cognition, and action in humans. We will focus our attention here not only because the technology and methods used in this research, and their strengths and weaknesses, are unique but also because this research has proven fascinating to the general public and policy-makers, which substantially broadens the impact that this research may have.

Ultimately, while the marriage of neuroscience and philosophy has tremendous potential to advance our understanding of ourselves as moral beings, we must be careful to limit both our scientific and philosophical conclusions to what the research can reasonably tell us, based on the limits both of the technology and of our ability to integrate centuries of philosophical thought into research design.

Introduction

Over the last 10–15 years, the tools of neuroscience have increasingly been employed to address questions long considered the realm of philosophers, including questions of moral sentiment, choice, cognition, and action. This research seeks to identify the structures within the brain that function in and are the basis of human morality, to improve our understanding, for example, of the role of emotion in moral decision-making, or the moral failings of psychopaths. If successful, such research might not only allow philosophers to refine their theories based on scientific facts but also might inform social policy.

While research on morality comes from a range of disciplines, including psychology, economics, and the social sciences, this chapter will focus on the subset of this research that uses the tools of neuroscience (primarily functional magnetic resonance imaging; fMRI) and is explicit about testing hypotheses related to moral sentiment, choice, cognition, and action in humans. We will focus our attention here not only because the technology and methods used in this research, and their strengths and weaknesses, are unique but also because this research has proven fascinating to the general public and policy-makers, which substantially broadens the impact that this research may have.

This fascination stems both from the technology used (fMRI) and from the questions being asked. The technology allows us to peer into the brain and watch it in action, which is both marvelous and mystifying: marvelous to have such access to the organ that is so closely linked to our identities; mystifying as we try to interpret what we see there.

It is important, however, to recognize the weaknesses of both the technology and of some efforts to solve philosophical puzzles through science. While the marriage of neuroscience and philosophy has tremendous potential to advance our understanding of ourselves as moral beings, we must be careful to limit both our scientific and philosophical conclusions to what the research can reasonably tell us, based on the limits both of the technology and of our ability to integrate centuries of philosophical thought into research design.

Morality

Throughout this chapter, the term "human morality" will be used rather narrowly, to refer to the attempt to live according to some view of what we should do or how we should live and to work out which such view, if any, can be justified and how our own view might be improved. It is not assumed that such views are normally (or ever) fully articulate or coherent, or that most people explicitly consider their views about how to live before deciding what to do. Furthermore, no particular stance is taken regarding how our views about how we should live and our conduct are related. That said, the discussion will be limited to the conduct of persons who are capable of formulating and rationally revising views about how to live and what to do and of trying to live by them.

The capacity for moral action, in this sense, requires a constellation of more specific capacities. A moral agent needs to be able to ask herself how she should live, to formulate answers to that question, and to rationally revise them as necessary. She will need to be able to experience those emotions that her moral views specifically require, for instance, empathy or outrage. She needs to recognize occasions when her moral views apply and to see what they require of her. And she needs the various capacities that make up the ability to regulate her own conduct. Neuroscientific research that illuminates any of these capacities helps us to understand ourselves as moral agents.

The State of the Science

The small but growing neuroscience literature addressing questions of moral sentiment, choice, cognition, and action may be divided into three broad and somewhat overlapping categories: the role of emotion in moral decision-making, neural correlates of other aspects of human morality, and research on the neural correlates of psychopathy. This section will briefly review this research and its findings concerning the neural correlates of moral sentiment, choice, cognition, and action.

The Role of Emotion

"Neuromorality" research was propelled into the spotlight with a study by Greene et al. (2001) investigating the role of emotion in moral decision-making. Much of this work has involved presenting research subjects with moral dilemmas and asking them to make judgments about the appropriate action in each case while undergoing fMRI. Greene's work, in particular, has been motivated by a pair of philosophical thought experiments called the "trolley" and "footbridge" dilemmas. These dilemmas are very similar but with one significant difference. In each scenario, five people are in the path of an oncoming trolley, and there is one way to save them. In the "trolley" case, five people can be saved by diverting the trolley onto another track, where it will run over one person rather than the original five. In the "footbridge" case, five people can be saved only by pushing a large man onto the tracks below, where he will stop the trolley before it reaches the five. In both cases, one person dies and five people are saved, and in both cases, it is our action that causes that one to die. Despite these similarities, most people respond quite differently to these two dilemmas: they think that it is permissible to divert the trolley, but not to push the man onto the track. In an attempt to tease out the apparent conundrum, Greene et al. posed a range of moral and nonmoral questions to their subjects and looked to see what their brains were doing while they made their decisions.

Greene et al. (2001) hypothesized that the difference in how we respond to these moral dilemmas hinges on the degree to which they engage our emotions. As the authors state, thinking about having to lay hands on a fellow human and push him to his death is "more emotionally salient than the thought of hitting a switch that will cause a trolley to produce similar consequences, and it is this emotional response that accounts for people's tendency to treat these cases differently." As such, the group proposed, when people respond to these "personal" moral dilemmas, they will be more likely to engage the emotional centers of the brain than when they respond to "impersonal" moral dilemmas, such as the trolley problem. In fact, they reported that in both of their experiments, parts of the brain implicated in emotion [medial portions of Brodmann's areas (BA) 9 and 10 (medial frontal gyrus), BA 31 (posterior cingulate gyrus), and BA 39 (angular gyrus, bilateral)] were significantly more active when responding to "moral-personal" scenarios than when responding to "moral-impersonal" or nonmoral scenarios.

Based on this work and subsequent investigations (e.g., Greene et al. 2004; Greene 2007a; Koenigs et al. 2007; Valdesolo and DeSteno 2006), Greene and colleagues affirmed a dual-process theory of moral decision-making (Greene 2007a, b). Greene and colleagues' version of the dual-process theory states that in cases like the footbridge dilemma, moral disapproval is driven by a prepotent negative emotional response involving the medial prefrontal cortex. In a parallel process, utilitarian moral reasoning is employed, potentially supported by the dorsolateral prefrontal cortex (DLPFC). In the absence of an overriding emotional response, "utilitarian reasoning prevails" (Greene 2007a). In difficult cases such as

the footbridge dilemma, the two processes (prepotent emotions and utilitarian reasoning) conflict, signaling the anterior cingulate cortex that cognitive control is needed, which is then implemented by the anterior DLPFC (Greene 2007a).

Moll and colleagues conducted a parallel research program looking at the role of emotion. Based on their and others' work (e.g., Ciaramelli et al. 2007; Koenigs et al. 2007; Moll et al. 2001, 2002a, b; Moll and de Oliveira-Souza 2007a, b; Moll and Schulkin 2009), they have concluded that moral judgment invokes brain regions including the frontopolar cortex (FPC, BA10), lateral and medial sectors of the orbitofrontal cortex (OFC, VMPFC, BA11), medial frontal gyrus (BA 10/46 and 9), right anterior temporal cortex (aTC), and the superior temporal sulcus (STS) of the left hemisphere. Further, Moll et al. propose that the association of moral emotions with ordinary social events or interactions may help motivate your response to such events in the future (Moll et al. 2002b).

Moll and colleagues view social sensibility as fundamental to morality and frame their research accordingly (Moll and Schulkin 2009). These researchers interpret the collective literature somewhat differently than Greene and colleagues, suggesting that rather than the dual-process theory in which reason and emotion are in competition, a more parsimonious explanation for the data is that moral sentiments emerge from the integration of cognition, emotion, and other aspects of decision-making. Moll and colleagues propose that the ability to experience "prosocial" moral sentiments, such as compassion and gratitude, is key, that this is facilitated by the VMPFC-FPC (Moll et al. 2005a; Moll and Schulkin 2009), and that such sentiments emerge from the integration of emotional and cognitive mechanisms, rather than from competition between the two (Moll et al. 2005b, 2006; Moll and de Oliveira-Souza 2007b). Further, they contend that while there is functional segregation at play in the brain, it is between prosocial sentiments and socially aversive sentiments (e.g., indignation, shame), which tend to activate different brain regions, rather than between emotion and cognition, as Greene and colleagues suggest (Moll and Schulkin 2009).

Beyond the Role of Emotion

While much neuromorality research has focused on the role of emotion in moral decision-making, there have also been a number of studies looking at specific moral emotions and other aspects of human morality. For example, Moll et al. (2005a) attempted to distinguish the neurological bases of nonmoral ("pure") disgust and moral disgust (defined as indignation) (Moll et al. 2005b). In this study, 14 subjects were observed with fMRI while reading neutral statements, statements designed to evoke pure disgust (e.g., "One night you were walking on a street. You saw a cat eating its own excrement"), and statements designed to evoke indignation (e.g., "As you arrived home, you saw that the nurse had put a spider on the baby's face"). They found that the patterns of brain activation invoked by pure disgust and indignation were highly similar, with only a few relevant differences, including the differential activation of the more anterior sectors of the OFC, left piriform

cortex (BA13, 38, 47), anterior superior frontal gyrus (SFG; BA8, 9), and right anterior inferior temporal gyrus (ITG; BA20, 21) in the experience of indignation versus pure disgust.

Knoch et al. (2006) investigated the underpinnings of reciprocal fairness in the context of the ultimatum game, in which one player offers a portion of her funds to the other player and the second player can accept or reject offers. If the second player accepts, she receives the agreed portion of the total. If the second player rejects, both players receive $0. While this game was being played, Knoch and colleagues disrupted the DLPFC using transcranial magnetic stimulation and found that stimulation/disruption of the right (but not the left) DLPFC lowered the probability that the second player would reject low offers, though she still judged those offers to be unfair. The authors concluded that the right DLPFC supports the implementation of fairness goals when faced with a conflict between fairness and self-interest.

Young and Saxe (2008) looked at the role of “theory of mind” in moral judgment. In order to judge a person’s actions, we must have beliefs about what that person believes she is doing and what she thinks the outcome of her action will be. For example, if a woman puts poison in her colleague’s coffee, reasonably thinking that the substance is sugar rather than poison, we will judge her actions differently than if we believe that she knows that the substance she is putting in her colleague’s coffee is poison. Young and Saxe reported that the process of belief attribution in such cases includes two components: belief encoding (forming a representation of our colleague’s belief about the white substance) and belief integration (using that belief in combination with the outcome to judge the morality of our colleague’s action). Belief encoding, they report, recruits the right temporoparietal junction (RTPJ), the precuneus (BA7), and, to a lesser extent, the left temporoparietal junction (LTPJ). The authors report that these same regions are recruited, though to a different extent, for belief integration, wherein the encoded beliefs are incorporated with other relevant information, such as the outcome of the action in question, enabling moral judgment. However, they conclude that the medial prefrontal cortex is uniquely recruited for processing belief valence (e.g., negative vs. neutral).

More recently, Cushman et al. (2012) examined the difference between people’s reactions to and judgments about harmful actions versus harmful omissions, as a way to get at the broader question of whether moral judgments tend to be intuitive and automatic (the “automaticity hypothesis”) or rather the result of explicit reasoning (the “rule hypothesis”). It has long been recognized that individuals tend to judge more harshly harmful actions than harmful omissions (the “omission effect”). The authors reported that their data support a version of the automaticity hypothesis and fail to support the rule hypothesis, demonstrating that increased cognition (evidenced by differential recruitment of the DLPFC; BA9, 10) was required not when individuals conformed to the omission effect but when they failed to demonstrate this bias. Put another way, Cushman et al. suggest that the omission effect is the result of automatic processes, but that these processes can be overcome with explicit cognitive effort.

Psychopathy

One distinct area of neuromorality research focuses on the differences between psychopathic and non-psychopathic individuals (Blair 2007; Glenn et al. 2009; Gregory 2012; Harenski et al. 2010; Raine and Yang 2006; Weber et al. 2008). There are two primary explanatory theories for psychopathy: the "somatic marker hypothesis" (Damasio 1994) and the "violence inhibition mechanism" model (Blair 1995). The somatic marker hypothesis holds that an inability to associate physiologic responses ("somatic markers"), such as increased heart rate, and related emotions, such as fear or anxiety, with stimuli (e.g., seeing a snake in your path) impairs the individual's ability to make efficient and appropriate decisions based on past experience. Such somatic markers can help us or bias us to choose one action over another based on how it made us feel the last time we faced a similar choice. The loss of this ability is associated with damage to the vmPFC and the resulting poor decision-making and antisocial behavior. The violence inhibition mechanism model holds that humans, like other social animals, have an innate response to "submissive" or "distress" cues from others, such as a look of fear or pain on someone's face, and that these cues prevent us from committing violence against those who are not aggressors. The loss of the violence inhibition mechanism leads to a lack of empathy and the development of psychopathy.

Several research teams have reported and discussed compromised function in the vmPFC and the amygdala (Blair 2007; Glenn et al. 2009; Harenski et al. 2010; Raine and Yang 2006). Other areas that have been implicated in psychopathy include the temporal cortices and posterior cingulate (e.g., Harenski et al. 2010; Raine and Yang 2006). This dysfunction, it is claimed, results in an inability to learn to associate particular actions with others' distress and to be aversive to those actions as a result.

Scientific Issues

As Poldrack has written (Poldrack 2006, 2007, 2011), there is still much that we do not understand about what fMRI can and cannot tell us. Poldrack and others have warned in particular of the related challenges of region of interest (ROI) analyses and reverse inference with fMRI data. ROI analyses, which focus preferentially on the results from specific regions of the brain following collection of data from the whole brain, are done for several reasons, including simply making the task more manageable by exploring the results for particular regions of the brain one at a time in data collected in a detailed whole-brain analysis, statistical analysis to reduce the degree of correction necessary for multiple testing by narrowing the amount of data being analyzed, and examining the activity of "functionally coherent" regions of the brain in response to a manipulation (Poldrack 2007). The last of these, the use of ROI analysis to assess the response of a particular brain region to an experimental manipulation, can be particularly

challenging due to the difficulty of attributing a single function of interest to a particular brain region (such "reverse inference" is discussed further below). This is compounded by the difficulty of defining ROIs themselves – not only is it a subjective process generally, but an ROI will be slightly different for each research participant due to structural and functional variability across human brains (Poldrack 2007). As such, multiple subjective judgments are being made in any given experiment, including about the threshold for calling a difference in regional activation between the case and control conditions, determining the size of the region of interest, and placing the demarcating line around the region of interest. For this reason, researchers must be circumspect in the conclusions drawn on the basis of the results.

Reverse inference in fMRI studies is the fairly common process of inferring the presence of particular mental states based on brain activation patterns in response to a manipulation. This process has been criticized for reasons including the lack of adequate background information (e.g., base rates of activation in particular brain regions), the lack of specificity of many brain regions, and the lack of clear or consistent ways of talking about the functions of different brain regions (Poldrack 2006, 2011). The lack of adequate background information, due in part to both the small sample sizes used in most fMRI studies and the historical lack of adequate mechanisms and incentives to share research scans with other researchers, combines with the inherently relative nature of fMRI measurements to make it difficult, if not impossible, to estimate base rates of activation for a ROI (however that has been defined in any given study), though this may be improving (Poldrack et al. 2011). Without knowing the base rate of activation, it is difficult to appreciate the strength of a reverse inference. Further, even if one does know the base rate of activation, if a ROI is activated in response to multiple stimuli (i.e., it lacks specificity), determining which mental state is present is difficult. Finally, even the different mental states that may be present can be difficult to define, since there is no standard way of talking about or describing them (Poldrack 2006, 2011). For all of those reasons, ascribing causation where a mere association is present should be done only with caution, and causal claims that are made should be described as tentative.

One of the reasons reverse inference can lead to inappropriate conclusions is that any given brain region may have multiple functions (Poldrack 2006). For example, in recent reports, the precuneus, which was found, as discussed above, to be involved in belief encoding (Young and Saxe 2008), has also been described as: "involved in determining the precise location of relevant stimuli from various sensory modalities as well as in the subject's preparation to act on it" (Zündorf et al. 2013); "a brain region known to be implicated in visuospatial imagery and more specifically, in shifting attention between targets" (Fonville et al. 2013); "implicated in higher-order cognitive functions, and might therefore contribute in accounting for the abnormal processing of [deontological guilt] stimuli in OCD patients" (Basile et al. 2013); "more involved in processing Chinese characters as compared with English words, more involved in Chinese adults than Chinese children during character reading, and more involved in Chinese

[as a second language] learners with a higher performance" (Cao et al. 2013); and "activations during task switching have been proposed to reflect attentional demands when updating stimulus–response associations" (Barber and Carter 2005). Thus, mere activation difference between a case and control state does not allow the researcher to draw a straight line between a particular mental state invoked by the research manipulation and, in this case, the precuneus.

While lesioning and other behavioral data can strengthen the case for a relationship between a brain area and a capacity, this is also not infallible. In the event of brain injury, other areas of the brain can take over a function previously supported by an injured area, thus compensating for the loss; for example, following a stroke which damages the part of the brain normally associated with spoken language, other parts of the brain can be recruited to support this function, allowing the patient to regain her ability to speak (e.g., Ohyama et al. 1996; Thiel et al. 2006). Given the known plasticity of the brain, as well as related structural and functional variability across individuals, even associations with lesioning or behavioral support should be considered carefully before the associations are claimed as evidence of causation.

Given how little we know, and related to the concerns raised above, there is also a worry about the degree to which the methods we use and the assumptions we make shape and constrain what we will find. This is of particular concern given the small numbers (often ten or fewer) of subjects in many of the studies published thus far investigating neural correlates of moral sentiment, choice, cognition, and action. In addition, these subjects are often university students (age 18–24 years) – a small and particular subset of the population who are unlikely either to be representative of the population as a whole or to capture humanity's diversity in brain structure and function and other relevant factors (e.g., Young and Saxe 2011). What if the results describe only how students at elite universities think about moral dilemmas?

Finally, there is the difficulty of the subject matter itself: moral sentiment, choice, cognition, and action. How one feels about or considers a particular moral dilemma may be colored by one's lived experience. What if contrived moral dilemmas like the "trolley" and "footbridge" scenarios, dilemmas unrelated to the lives of these students, recruit different neural regions or networks than the mundane moral problems of their lives (e.g., Should I cheat on this test?)? Could these actually be studies of moral imagination? To some degree this is unavoidable, as it would be unethical to put subjects in actual moral dilemmas, but again, this should caution against too strong conclusions. As noted by Moll and de Oliveira-Souza (2007b) in discussing the results of a particular study, "That VMPFC patients make more prosocial choices (from a utilitarian perspective) is a reminder of the gulf that divides observable behaviors and internal motivations." To better understand such complex questions of human morality, further research on the basic capacities underlying moral sentiment, choice, cognition, and action – such as decision-making and acting on one's decisions – will be required. Such understanding is necessary for developing testable hypotheses about how these capacities and the mechanisms that enable them might be involved in moral sentiment, choice, cognition, and action.

Philosophical Issues

Neuroscientists who study morality are concerned with questions about how we actually make moral decisions. Moral philosophers, on the other hand, are concerned with what we ought to do: with which decisions are the right ones, which kinds of lives we ought to lead, and which goals we ought to pursue. These are quite different questions, and while answers to the first set of questions are surely relevant to the second, they are not themselves answers to them, nor do they imply such answers in any obvious way. In this section we will consider some ways in which neuroscientific research about morality might be thought to bear on ethics.

Settling Questions About the Basis of Morality

Some neuroscientists have taken their work to bear directly on fundamental issues in moral philosophy. Thus, Marc Hauser writes: "Neuroimaging is at its best and most useful when there are competing psychological theories, with one side proposing a unitary mechanism and the other proposing multiple mechanisms. If you are a Kantian creature, you think that the key process underlying moral judgment is deliberate reasoning based on clearly articulated principles. Consequently, the areas involved in such reasoning, be they specific to morality or not, should not only turn on but have the most active voice. If you are a Humean creature, you think that only our emotions play a role in moral judgments, and, thus, the circuitry underlying the production and perception of emotions should turn on. There are, of course, many other possibilities . . ." (Hauser 2006, p. 220).

In this passage and elsewhere, Hauser confuses two sets of questions. The first set, which neuroimaging can illuminate, is: what brain structures are activated when we make moral choices in everyday life? Specifically, what roles do reason and emotion play in these moral choices? The second set, on which the bearing of neuroscience is much less clear, is: How might we justify moral claims and what role might reason and emotion play in that justification? These two sets of questions are quite different, and answers to one do not necessarily translate into answers to the other.

The moral philosophers Hauser mentions are not describing the ways in which we make moral decisions in everyday life but the basis on which moral claims can be justified. Hume believes that we cannot justify any moral claims on the basis of reason alone; Kant believes that we can. Neither view directly implies anything about how we make moral judgments in everyday life.

Morality as Rationalization of Choices Made on Other Grounds

It has been suggested (Funk and Gazzaniga 2009; Haidt 2001) that moral reasoning is simply a post hoc rationalization for decisions made on nonrational grounds. If this were true, then the questions that interest moral philosophers would have no

practical point: if our views on moral questions could not affect our conduct, then our interest in answering those questions correctly would be purely academic. It would not be surprising to discover that people sometimes believe that they have acted on a moral principle when in fact their actions had a different cause. People are sometimes mistaken and they sometimes confabulate, as Funk and Gazzaniga note.

However, it would be quite surprising to discover that we never ask ourselves what to do, try to reason our way to an answer, and then act on it. It is worth noting both how difficult it would be to provide evidence for this claim, which purports to hold for all moral reasoning, and how implausible it is, since the fact that we sometimes do ask what we should do and then act on our answer seems clear from experience. Moreover, it is not clear that moral reasoning could always be confabulation, since we sometimes engage in moral reasoning before making a decision, rather than producing it as a rationalization after the fact.

In fact, Haidt does not deny that we sometimes reason our way to a decision and then act on it. He just thinks that we do so only "rarely" (Haidt 2001, p. 189). Haidt presents this view as a hypothesis in need of further testing, and his arguments for it are directed not against the view that our moral views can influence what we do but against the view that virtually nothing other than reason governs moral action, a view that his opponents need not accept. Similarly, Funk and Gazzaniga "suggest" that "rather than a causal determinant in the moral decision-making process, moral reasoning is most usefully thought of as an attempt to explain the cause and effect of our moral intuitions that draws upon all available explicit information about a given situation" (2009, p. 680). While they do provide evidence that reasoning is sometimes post hoc rationalization or confabulation rather than a cause of our action, they do not provide evidence that this is always true. Moreover, the evidence they do provide is suggestive rather than conclusive.

Greene's Dual-Process Theory

As noted earlier, Greene proposes a dual-process theory of moral reasoning. For evolutionary reasons, he argues, we are normally reluctant to hurt other people. In "personal" moral dilemmas in which we must decide whether or not to hurt another person, our moral judgments tend to be driven by these emotional responses. But in "impersonal" cases in which the harm we impose on others is less obvious, those emotions do not come into play, and we can rely instead on a rational assessment of the costs and benefits of our alternatives.

Greene argues (Greene 2007b) that this view provides evidence that deontological moral theories are rationally unfounded. "There is a remarkable correspondence between what rationalist deontological theories tell us to do and what our emotions tell us to do" (Greene 2007b, p. 68). What explains this correspondence, Greene argues, is not that our emotions track an independent and rationally

discoverable moral truth but that many of our moral beliefs, and in particular those that we draw on in formulating deontological views, are essentially rationalizations of our emotional responses.

One might question this argument on a number of grounds. Greene sometimes seems too quick to draw general lessons from the contrast between the "trolley" and "footbridge" cases. In those cases, unwillingness to directly hurt others lines up with deontology, and setting our emotional responses aside lines up with consequentialism. But this is not always true: when we need to choose between lying and telling a truth that will both hurt someone directly and produce bad consequences overall, it is consequentialism that lines up with our emotional responses, and deontology that requires that we set those emotional responses to one side on the basis of rational reflection.

Similarly, Greene's claim that there is a "remarkable correspondence" between the views of rationalist deontologists and our instinctive aversion to harming others relies in part on his choice of examples. There are cases, some of which Greene cites, in which our moral intuitions fit deontological moral theory. But there are other cases in which they do not. Kant believed that it is wrong to lie to a would-be murderer who asks where he can find an innocent person whom he wants to kill, and that we should do what is right regardless of the consequences. Neither claim accords with most people's moral intuitions.

The most important problem with Greene's argument, however, is that it begs the question against the rationalist deontologist. Rationalist deontologists believe that we can justify some set of non-consequentialist moral rules on the basis of moral reasoning. If they can provide such a justification – a valid argument for the truth of those moral rules – then that argument would explain why they accepted those rules, and any correspondence between those rules and our emotional responses to hurting other people would be coincidental. If there is no such justification, then rationalist deontology is false, whatever its relation to our instinctive emotional responses.

Greene's dual-process view of moral reasoning does provide a possible explanation for some of our non-consequentialist moral judgments. In so doing it helps us to understand how our moral lives might be as they are even if we have no rational justification for some of our moral beliefs. But his theory does not imply that there is no such justification.

What Neuroscience Can Do

Morality is concerned with questions about how should we live. Neuroscience does not aim to answer these questions. For this reason, neuroscientific results are unlikely to directly imply the truth or falsity of any general moral view. However, precisely because neuroscience tells us about the brain structures that underlie our ordinary moral decisions, our attempts at self-governance, and our conduct, they are relevant to moral philosophy in a different way.

Morality is by its nature practical. We try to figure out how we ought to live not only because this question is interesting but because we want to put our answers into practice. The more we understand the mechanisms that underlie our conduct, the more clearly we can see how these mechanisms might go wrong and how to make it as likely as possible that they function as we would wish them to.

For instance, Greene and Paxton (2009) attempted to determine whether subjects who did not take advantage of opportunities for dishonest gain acted as they did because they resisted temptation or because they did not find those opportunities tempting in the first place. Greene and Paxton argued that if subjects behaved honestly because they were able to resist temptation, then they ought to display increased activity in brain regions associated with self-regulation and control, whereas if they behaved honestly because they were not tempted by the prospect of dishonest gain, such activity would not be required. They reported that subjects who behaved honestly displayed no additional activity in regions associated with self-control and therefore concluded that those subjects generally were not tempted by dishonest gain. Moreover, they found that subjects who behaved dishonestly did display increased activity in those regions, and that this was true even on those occasions when those subjects acted honestly.

Greene and Paxton's results do not imply that people who want to be honest have no reason to try to develop willpower. After all, if one finds oneself tempted by dishonest gains, then it is too late to try to make oneself into the sort of person who would not be so tempted. But they do imply that if we can manage to avoid being tempted by dishonesty, either by cultivating better habits, avoiding occasions for dishonesty, reframing the issues, or by some other means, then we would be well advised to do so. Moreover, their results imply that it is more important to try to avoid being tempted in the first place than to develop willpower. Had Greene and Paxton found the opposite – that temptation affects everyone equally, but only some are able to resist it – then the opposite would be true.

These results do not themselves settle any major debates in moral philosophy. Arguably, no plausible results in neuroscience could settle those debates. Instead, they help us to understand what is involved in actually trying to live a moral life and in so doing illuminate the terrain moral agents must navigate, the obstacles they are likely to confront, and the tools their brains provide to help them on the way.

Research Ethics

There is a great deal we can learn from the brain with current technology, but scientific integrity demands honesty and modesty in the face of many unknowns. Investigators must be careful to pose research questions that can be rigorously addressed given the current state of the technology, the outcomes we can reliably measure, and our ability to analyze and understand the resulting data. Investigators must also be appropriately circumspect in drawing conclusions from this work. As the field develops, perhaps more will be able to be said about the relationship of events in the brain to the capacity to make moral decisions, but currently, justifiable

claims based on this research are fairly modest. Caution may be particularly important in this field given the strong public interest in questions of moral sentiments and the brain. Further, the answers to such questions may have direct implications for our understanding of moral responsibility and ultimately criminal responsibility, which also suggests caution against making too-broad claims based on current research.

Beyond these issues of scientific integrity, so long as the science is focused on imaging studies, there are few other unique research ethics issues raised, except perhaps in the case of psychopaths who may be incarcerated or otherwise unable to give fully informed and voluntary consent. If or when the field begins to venture into research on enhancing (or diminishing) the morality of research subjects, the ethical landscape will surely change.

Conclusion and Future Directions

While neuromorality research is interesting, has great promise to inform many questions related to the neural correlates of moral sentiment, choice, cognition and action, and has undeniably captured the imagination of the public, this is a young and in many ways primitive field. Both neuromorality researchers and the public must appreciate and acknowledge the difference between determining the neural correlates of moral sentiment, choice, cognition, and action in any particular case and resolving the difficult philosophical questions of what is or how to lead a moral life. Going forward, additional collaboration between philosophers and neuroscientists is critical to improving the chances that the research has the desired effect of informing (though not solving) debates in moral philosophy and at the same time that researchers do not overstate the philosophical implications of their conclusions.

Cross-References

▶ Beyond Dual-processes: The Interplay of Reason and Emotion in Moral Judgment
▶ Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
▶ Historical and Ethical Perspectives of Modern Neuroimaging
▶ Human Brain Research and Ethics
▶ Moral Cognition: Introduction
▶ Moral Intuition in Philosophy and Psychology
▶ Neuroimaging Neuroethics: Introduction
▶ The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy
▶ The Morality of Moral Neuroenhancement
▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering

References

Barber, A. D., & Carter, C. S. (2005). Cognitive control involved in overcoming prepotent response tendencies and switching between tasks. *Cerebral Cortex, 15*, 899–912.

Basile, B., Mancini, F., Macaluso, E., Caltagirone, C., & Bozzali, M. (2013). Abnormal processing of deontological guilt in obsessive-compulsive disorder. *Brain Structure and Function*, May 17 [Epub ahead of print].

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition, 57*, 1–29.

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences, 11*, 387–392 (Regul Ed).

Cao, F., Tao, R., Liu, L., Perfetti, C. A., & Booth, J. R. (2013). High proficiency in a second language is characterized by greater involvement of the first language network: Evidence from Chinese learners of English. *Journal of Cognitive Neuroscience, 25*, 1649–1663.

Ciaramelli, E., Muccioli, M., Ladavas, E., & Pellegrino, G. D. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2*, 84–92.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2012). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience, 7*, 888–895.

Damasio, A. R. (1994). *Descartes' error*. New York: Putnam.

Fonville, L., Lao-Kaim, N. P., Giampietro, V., Van den Eynde, F., Davies, H., Lounes, N., Andrew, C., Dalton, J., Simmons, A., Williams, S. C. R., Baron-Cohen, S., & Tchanturia, K. (2013). Evaluation of enhanced attention to local detail in anorexia nervosa using the embedded figures test; an fMRI study. *PLoS One, 8*(5), e63964.

Funk, C. M., & Gazzaniga, M. S. (2009). The functional brain architecture of human morality. *Current Opinion in Neurobiology, 19*, 678–681.

Glenn, A. L., Raine, A., & Schug, R. A. (2009). The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry, 14*, 5–6.

Greene, J. D. (2007a). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11*, 322–323 (Regul Ed).

Greene, J. D. (2007b). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 3). Cambridge: The MIT Press.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences, 106*, 12506–12511.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Gregory, S. (2012). The antisocial brain: Psychopathy matters, a structural MRI investigation of antisocial male violent offenders. *Archives of General Psychiatry, 69*, 962–972.

Haidt, J. (2001). The emotional dog and its rationalist tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Harenski, C. L., Harenski, K. A., Shane, M. S., & Kiehl, K. A. (2010). Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology, 119*, 863–874.

Hauser, M. (2006). *Moral minds*. New York: HarperCollins.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science, 314*, 829–832.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*, 908–911.

Moll, J., & Oliveira-Souza, R. (2007a). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences, 11*, 319–321 (Regul Ed).

Moll, J., & de Oliveira-Souza, R. (2007b). Response to greene: Moral sentiments and reason: Friends or foes? *Trends in Cognitive Sciences, 11*, 323–324 (Regul Ed).
Moll, J., & Schulkin, J. (2009). Social attachment and aversion in human moral cognition. *Neuroscience and Biobehavioral Reviews, 33*, 456–465.
Moll, J., Eslinger, P. J., & Oliveira-Souza, R. (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task: Preliminary functional MRI results in normal subjects. *Arquivos de Neuro-Psiquiatria, 59*, 657–664.
Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002a). Functional networks in emotional moral and nonmoral social judgments. *NeuroImage, 16*, 696–703.
Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002b). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience, 22*, 2730–2736.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005a). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*, 799–809.
Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignácio, F. A., Bramati, I. E., Caparelli-Dáquer, E. M., & Eslinger, P. J. (2005b). The moral affiliations of disgust: A functional MRI study. *Cognitive and Behavioral Neurology, 18*, 68–78.
Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 15623–15628.
Ohyama, M., Senda, M., Kitamura, S., Ishii, K., Mishina, M., & Terashi, A. (1996). Role of the nondominant hemisphere and undamaged area during word repetition in poststroke aphasics. A PET activation study. *Stroke, 27*, 897–903.
Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*, 59–63 (Regul Ed).
Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience, 2*, 67–70.
Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron, 72*, 692–697.
Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D. S., Sabb, F. W., & Bilder, R. M. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics, 5*, 17.
Raine, A., & Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. *Social Cognitive and Affective Neuroscience, 1*, 203–213.
Thiel, A., Habedank, B., Herholz, K., Kessler, J., Winhuisen, L., Haupt, W. F., & Heiss, W. D. (2006). From the left to the right: How the brain compensates progressive loss of language function. *Brain and Language, 98*(1), 57–65.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*, 476–477.
Weber, S., Habel, U., Amunts, K., & Schneider, F. (2008). Structural brain abnormalities in psychopaths – A review. *Behavioral Sciences & the Law, 26*, 7–28.
Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*, 1912–1920.
Young, L., & Saxe, R. (2011). Moral universals and individual differences. *Emotion Review, 3*, 323–324.
Zündorf, I. C., Lewald, J., & Karnath, H. O. (2013). Neural correlates of sound localization in complex acoustic environments. *PLoS One, 8*(5), e64259.

Section XVI

Neuroenhancement

Neuroenhancement

75

Bert Gordijn

Contents

Abstract

This chapter is an introduction to the section on neuroenhancement, which is regarded as an intervention in the central nervous system in order to "improve" certain aspects of a person's "healthy" or "normal" performance. This can be accomplished by pharmaceutical means, surgery, and/or technology. The section on neuroenhancement begins with a chapter looking at pharmaceutical mood enhancement. It then centers on smart drugs followed by a chapter on the use of brain-computer interfaces for enhancement purposes and another one analyzing moral neuroenhancement. The section concludes with a chapter presenting some general reflections on neuroenhancement.

Introduction

Since the dawn of history, people have aimed at self-improvement. They have attempted to develop their traits and faculties in a variety of ways depending on the specific set of cultural ideals *en vogue* at the time. Gilgamesh searched for immortality. The Homeric heroes strived for excellence in fights and wars. The Olympic

B. Gordijn
Institute of Ethics, Dublin City University, Dublin, Ireland
e-mail: bert.gordijn@dcu.ie

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_88,

athletes pushed the boundaries in sports. Early Christian hermits aspired after spirituality and holiness. Medieval manuscript illuminators arduously developed their artistic skills. Currently, the "self-improvement industry" is huge.

For a long time, medicine did not play any role of significance in this widespread human attempt at self-improvement. Hippocratic medicine was focused on trying to make sick people healthy again. It conceptualized health as equilibrium between four humors: blood, phlegm, black, and yellow bile. An imbalance of these humors caused disease. Therefore, Hippocratic medicine was led by the normative ideal of *restitutio ad integrum*: restoration of the original balance. The Hippocratic conceptualization of health and disease, and the normative framework associated with it, dominated medicine for more than two thousand years.

In addition to these Hippocratic notions of curing diseases and restoring health, ideas of using medicine in order to improve people beyond what could be regarded as healthy or normal were first tentatively addressed by Francis Bacon in his *New Atlantis* (1627) and Rene Descartes in his *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences* (1637). Both philosophers focused on an imagined and desired future state of medicine. They both argued that medicine would have a huge potential for future development, if only it employed the right scientific methodology. A century and a half later, these early ideas were even more explicitly heralded by Marquis de Condorcet in *Esquisse d'un tableau historique des progrès de l'esprit humain* (1795), where he discussed the extension of the maximum lifespan and the limitless perfectibility of man.[1] These seventeenth and eighteenth century enthusiastic ideas about medicine had increasingly moved away from the traditional Hippocratic normative framework. However, focused as they were on an anticipated future state of medical science, they were somewhat remote from medical practice at the coalface.[2]

Arguably, the state of medicine imagined in the writings of Bacon, Descartes, and Condorcet has now at least partially been realized. It is true, we are still neither able to eradicate all pain and suffering, nor to extend the maximum human life span. Nevertheless, medical enhancement is a growth industry. A large part of this industry is currently focused on the improvement of looks. Modern medicine and dentistry have advanced countless new approaches to meet this goal: botox, breast implants, laser hair removal, liposuction, face lifts, tooth reshaping, veneers, and tooth bleaching. Another important chunk is to do with the improvement of performance in sports and the military.

Besides aesthetics, athletic endurance, and military performance, expectations are currently also directed to the medical improvement of various aspects of our mental live. To this effect, two approaches have dominated the academic debate in the last few decades. First, the focus was on genetic interventions. A theoretical advantage of genetic enhancement was that improvements in the germ line would

[1]See Gordijn 2006 for further analysis of this early philosophical literature on the future prospects of medical science.

[2]See Wiesing 2008 for further analysis of the history of medical enhancement.

be inherited by future generations. As a result, improvements would not have to start from scratch in each new generation. Of course, this advantage would directly turn into a disadvantage in case disproportionate harms are passed on to future generations as a result of unsuccessful germ-line interventions. Be that as it may, the discussion around projections centered on germ-line enhancement were somewhat dampened when, in the course of the 1990s, clinical trials demonstrated that even "simple" somatic gene therapy turned out to be much more problematic than imagined in advance. The problems surrounding genetic interventions may have created more room for excited projections centered on neuroenhancement: medical interventions more directly targeting a healthy central nervous system in order to improve its functions. This shift has triggered a new debate on the prospects and the ethics of neuroenhancement.

In Google Scholar, publications with the word "neuroenhancement," in the title turned up for the first time in 2004 (search performed on 31 March 2013). As is to be expected, the term "neuroenhancement" is differently defined by different authors. For example, neuroenhancement is sometimes narrowed down to "the use of pharmaceuticals to enhance cognitive function in cognitively normal people" (Boot et al. 2012, p. 181; cf. Eickenhorst et al. 2012; Normann and Berger 2008; Partridge et al. 2011). Others use a broader concept of neuroenhancement with a more inclusive range of methods as well as additional mental functions as potential targets of improvement. An example of such a broader definition of neuroenhancement is the "improvement in the cognitive, emotional and motivational functions of healthy individuals through, inter alia, the use of drugs (Repantis et al. 2010, p. 187).

In this section, neuroenhancement is loosely defined in a broad sense as an intervention in the central nervous system, by using pharmaceutical means, surgery, and/or technology (brain-computer interfaces or other neurotechnologies), in order to "improve" certain aspects of its "healthy" or "normal" performance. Of course, there are meticulous debates about what exactly constitutes "health," what level of performance can be regarded as "normal," and what kind of change may be conceived of as "improvement." However, this section does not primarily focus on these basic philosophical questions, even if they cannot be avoided altogether. Instead, it first and foremost discusses a few important types of neuroenhancement and a selection of ethical questions that they raise, without attempting to present an exhaustive overview.

Section Overview

The neuroenhancement section starts with a chapter on ethical issues of pharmacological mood enhancement by Maartje Schermer. First, Schermer discusses the concept of mood enhancement. This concept turns out to be quite difficult to demarcate. In addition, she gives an historical overview of the ethical debate about mood enhancement. The early mood enhancement discussions start with the publication of Peter Kramer's book *Listening to Prozac* (1993) and center on

Selective Serotonine Reuptake Inhibitors (SSRIs), a set of drugs – Prozac being a well-known example – originally developed in the 1970s to tackle depression. These drugs were seemingly without any serious side effects and became hugely popular as mood enhancers. However, in the early 2000s, critical questions around the occurrence of side effects and the lack of efficacy surfaced. As a result, the popularity of SSRIs waned during the 2000s, shifting the debate from pharmacological mood enhancement to other kinds of neuroenhancement. However, Schermer believes it is very well possible that emerging neuromodulation technologies might rekindle the debate about mood enhancement. In her analysis of the ethical and broader philosophical issues triggered by the prospect of mood enhancement, she covers issues such as the real meaning of happiness, the significance of feelings of alienation, authenticity and its relation to self-discovery and self-creation, the use of psychotherapy versus pharmaceuticals in order to alter mood or character traits, as well as issues of medicalization, disease mongering, and the goals of medicine.

In the next chapter, Alena Buyx focuses on the ethical debate about smart drugs, which is quite intensive at present both in the scholarly and the public arena. Seeking to avoid that her review becomes too speculative, she first surveys the empirical evidence on smart drugs, looking in particular at data about the frequency of their use as well as their effectiveness. In order to delimit her review, she especially focuses on modafinil, methylphenidate, and amphetamines. These drugs have originally been developed for therapeutic purposes. However, they are now also being used off-label for the improvement of cognitive functions by people showing none of the clinical defects that these drugs were originally intended to address. In her review of the ethical debate, Buyx first zeros in on conceptual issues such as the distinction between therapy and enhancement, and what is natural and unnatural. Next, she discusses ethical issues on an individual level such as considerations around safety, efficacy, autonomy, and authenticity. Finally, Buyx analyzes ethical issues triggered by smart drug that might surface on a societal level, for example, increasing inequality and other detrimental societal effects like medicalization and disengagement.

Fiachra O'Brolcháin and Bert Gordijn discuss the use of brain-computer interfaces (BCIs) for enhancement purposes. BCIs pick up brain signals and translate them into commands directing an external gadget. Like the smart drugs and mood enhancers discussed above, BCIs have originally been developed for therapeutic purposes in order to enable patients with brutal physical impairments to control a computer and other devices "by the power of their thoughts." However, they are now increasingly being used for enhancement purposes, for example, for gaming or neurofeedback in sports and the arts. The shift from use for therapeutic to enhancement applications means that the target group of users, formerly exclusively severely injured patients, might slowly but surely become very big, potentially turning BCIs into a phenomenon with considerable societal effects. The chapter anticipates a potential future scenario of widespread BCI use for the purpose of enhancement. After a brief sketch of a variety of ethical issues, the discussion centers on concerns around privacy and autonomy. Further ethical debate is

imperative in order to facilitate regulation that steers the development and use of BCI technologies in desirable directions, reaping the opportunities and effectively tackling the challenges.

Next Thomas Douglas discusses moral neuroenhancement. In contrast to moral enhancement, which has a broad set of approaches at its disposal in order to increase morality, the narrower concept of moral neuroenhancement covers exclusively direct interventions in the brain and its operations, for example, by means of drugs or neuromodulation technologies. Douglas' chapter covers two kinds of moral neuroenhancement that have prevailed in recent scholarly debates, i.e., neuroenhancement of the moral status of individuals and of the moral desirability of their character, motivation, and behavior. Douglas mainly focuses on ethical issues around moral neuroenhancement only turning to questions of feasibility when relevant for the ethical debate. In doing so, he presents a detailed overview of the main lines of argument in this burgeoning academic debate on the ethics of moral neuroenhancement.

Walter Glannon concludes the neuroenhancement section with some general reflections on the topic. He focuses on cognitive and moral neuroenhancement, intentionally leaving out mood enhancement. This focus is due to the fact that – as already mentioned above – in the last couple of years, empirical studies have advanced somewhat disappointing results about the efficacy of mood enhancers when used by healthy people. Glannon is especially interested in assessing the argument stating neuroenhancement is wrong because it negatively modifies human nature. Glannon also discusses the risks and benefits of cognitive neuroenhancement concluding that long-term empirical studies are required to gain more precise knowledge about its benefit-risk ratio. In addition, he examines some of the potential societal effects of cognitive neuroenhancement, such as increased inequality and social coercion to use cognitive neuroenhancers. Finally, Glannon analyzes questions around empirical evidence of the effectiveness of moral neuroenhancement to improve moral sensitivity as well as issues concerning individuals' freedom not to enhance their own moral traits – which may under certain circumstances conflict with community interests if, for example, moral neuroenhancement turns out to be cheap, effective, and without side effects.

Conclusion

The chapters in this section demonstrate that there is a burgeoning academic debate about neuroenhancement. While this debate is intellectually interesting in its own right, it remains to be seen whether the future prospects of neuroenhancement, that inform the contemporary discussion, will materialize as currently imagined. Analogous to the bubbles and busts that seem to dominate economic cycles, ethical debates about emerging technologies do sometimes appear to be characterized by an early focus on somewhat grandiose prospects that are later debunked on the basis of increased empirical data. The debate about mood enhancement seems to be a good case in point.

So at this early stage in the development of neuroenhancement technologies, it is still unclear whether effective and safe ways of neuroenhancement will attain widespread application any time soon – or ever for that matter. And yet precisely because it is impossible to predict the future, it is prudent to be prepared for all eventualities. Of course, ethical discussions should be founded on the results of empirical research. But in order to anticipate upcoming developments, future scenarios have to be analyzed as well. In doing so, it can sometimes admittedly be difficult to avoid a focus on speculative prospects, if only because it is so much more interesting to analyze the ethics of a future scenario significantly different to our current condition than it is to scrutinize the ethical aspects of imminent incremental change. However, as long as the academic debate keenly adjusts its findings on the basis of newly available empirical information about emerging versions of neuroenhancement technologies, their traits, and the various effects of their employment, this should not be an insurmountable problem.

Cross-References

▶ Attention Deficit Hyperactivity Disorder: Improving Performance Through Brain–Computer Interface
▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethics of Pharmacological Mood Enhancement
▶ History of Neuroscience and Neuroethics: Introduction
▶ Impact of Brain Interventions on Personal Identity
▶ Mind Reading, Lie Detection, and Privacy
▶ Reflections on Neuroenhancement
▶ Research in Neuroenhancement
▶ Sensory Enhancement
▶ Smart Drugs: Ethical Issues
▶ The Morality of Moral Neuroenhancement
▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Bacon, F. (1974 [1627]). The new *Atlantis*. In A. Johnston (Ed.), *The advancement of learning and new Atlantis* (pp. 1–212). Oxford: Clarendon.

Boot, B. P., Partridge, B., & Hall, W. (2012). Letter to the editor: Better evidence for safety and efficacy is needed before neurologists prescribe drugs for neuroenhancement to healthy people. *Neurocase, 18*(3), 181–184.

de Condorcet, N. (1988 [1795]). *Esquisse* d'un tableau historique des progrès de l'esprit humain. In A. Pons (Ed.), *Esquisse d'un tableau historique des progrès de l'esprit humain suivi de Fragment sur l'Atlantide* (pp. 79–296). Paris: Flammarion.

Descartes, R. (1897–1913[1637]). *Discours* de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences. In C. Adam & P. Tannery (Eds.), *Oevres de Descartes* (13 vols.). Paris: Léopold Cerf, Band 6.

Eickenhorst, P., Vitzthum, K., Klapp, B. F., Groneberg, D., & Mache, S. (2012). Neuroenhancement among German University students: Motives, expectations, and relationship with psychoactive lifestyle drugs. *Journal of Psychoactive Drugs, 44*(5), 418–427.

Gordijn, B. (2006). *Medical utopias. Ethical reflections about emerging medical technologies*. Leuven/Paris/Dudley: Peeters.

Kramer, P. D. (1993). *Listening to prozac: A psychiatrist explores antidepressant drugs and the remaking of the self*. New York: Viking.

Normann, C., & Berger, M. (2008). Neuroenhancement: Status quo and perspectives. *European Archives of Psychiatry and Clinical Neuroscience, 258*, 110–114.

Partridge, B. J., Bell, S. K., Lucke, J. C., Yeates, S., & Hall, W. D. (2011). Smart drugs "as common as coffee": Media hype about neuroenhancement. *PloS One, 6*(11), e28416.

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*(3), 187.

Wiesing, U. (2008). The history of medical enhancement: From restitutio ad integrum to transformatio ad optimum? In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Posthumanity* (pp. 9–24). New York: Springer.

Ethics of Pharmacological Mood Enhancement

76

Maartje Schermer

Contents

Abstract

Pharmacological mood enhancement – the improvement of mood and related mental functions by means of pharmaceuticals – raises a number of philosophical, ethical, and social questions. These questions are partly in line with questions known from the broader debate on human enhancement and partly relate to the ways in which neuroscience and neurotechnologies are affecting our ideas about who we are and how we should understand ourselves. This contribution gives an overview of the mood-enhancement discussion as it has played out in the academic literature over the last 20 years.

M. Schermer
Department Medical Ethics and Philosophy of Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands
e-mail: m.schermer@erasmusmc.nl

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_90,

After a brief historical overview of mood enhancement and the associated ethical debate, the most important ethical and philosophical issues regarding mood enhancement are addressed in more detail. These include, first, the nature and value of happiness and human well-being. Secondly, issues regarding authenticity, identity, and the proper view of and attitude towards the self are discussed. Finally, there are questions about the proper role of medicine and the critique on the medicalization of psychological states and traits and the role of industry.

It is concluded that with the advent of new technologies and the increasing importance of optimal mental functioning in modern societies, these issues will remain important in the near future, even if they may no longer be grouped together under the heading of "mood enhancement."

Introduction

Pharmacological mood enhancement – the improvement of mood and related mental functions by means of pharmaceuticals – raises a number of philosophical, ethical, and social questions that have been discussed in both popular and academic literature over the past decades. On the one hand, these discussions fit into the more general human enhancement debate; on the other hand they also relate to the ways in which neuroscience and neurotechnologies are affecting our ideas about who we are and how we should understand ourselves. This contribution will give an overview of the discussion as it has played out in the academic bioethical and neuroethical literature over the last 20 years. It will present the main questions and concerns that have been raised and the different points of view taken on these issues and show how these are related to philosophical and ethical debates and positions on, for example, personal identity, human agency, the goals of medicine, or human well-being.

First, the concept of mood enhancement will be discussed, as well as a number of techniques that may be used for mood-enhancement purposes. Next, a brief historical overview of mood enhancement and the ethical debate on mood enhancement will be given. After that the most important ethical and philosophical issues regarding mood enhancement as they have come up in the mood-enhancement debate will be addressed in more detail. These can roughly be clustered in three groups: first, questions concerning the nature of happiness and well-being; second, questions concerning the effects of mood enhancement on the authenticity, personal identity, or self-image of the person using it; and third, social questions concerning the goals of medicine, the role of industry, and the medicalization of unhappiness.

What Does "Mood Enhancement" Mean?

Mood enhancement is a very slippery concept and is actually not used as such by many of the authors who have participated in the discussion. A first reason why

"mood enhancement" is difficult to define is due to the fact that it is not clear how "enhancement" should be defined. This has already raised significant debate in the literature. Some definitions clearly distinguish between treatment or therapy on the one hand and enhancement on the other. The definition by Juengst, for example, states that "enhancements are interventions designed to improve human form or functioning beyond what it takes to restore and sustain good health" (Juengst 1998, p. 29). A problem with this type of definition is that it depends on concepts of health and disease that are contested themselves; it presupposes a clear distinction between sickness and health that is difficult to maintain, especially in the field of psychiatry and psychology.

The so-called welfarist definition of enhancement does not face such problems, as it states that an enhancement is "any change in the biology or psychology of a person which increases the chance of leading a good life in the relevant set of circumstances" (Savulescu et al. 2011, p. 7). This includes both medical treatment and enhancement of human traits or functioning above "normal" levels, thus including both the treatment of depression or anxiety disorders and making people feel "better than well."

Finally, some definitions also stress the technological or biomedical means of enhancement interventions: "a biomedical enhancement is a deliberate intervention, applying biomedical science, which aims to improve an existing capacity that most or all human beings typically have [. . .] by acting on the body or the brain" (Buchanan 2011, p. 23). Non-technological means for mood enhancement, like cognitive therapy, physical exercise, or good company, are thereby excluded from the ethical debate.

A second reason why it is difficult to define what is meant by "mood enhancement" is that what is often discussed under that heading is not always clearly the improvement of *mood*. Remarkably, in the debate over the ethics of mood enhancement, little has been written about what "mood" exactly is, perhaps with the exception of a phenomenological analysis of mood by Svenaeus (2007). "Mood" tends to be equated with "feeling good" or "being happy," but it is sometimes also taken to refer to a wider variety of emotional states or emotional responses. Moreover, much of the debate is not about mood in a strict sense but about the enhancement of certain desirable character traits such as being spontaneous, outgoing, self-confident, and not shy. Since the 2003 report by President Bush's Council on Bioethics, *Beyond Therapy*, the use of pharmaceuticals to "erase bad memories" – as the use of beta-blockers for the treatment of post-traumatic stress disorder PTSS is called in that report – is also often classified as a case of mood enhancement.

A final specification is that mood enhancement mainly refers to specific means of mood enhancement. In a sense, mood enhancement is nothing new because humans have always used herbal mixtures, drugs, or alcohol to enhance mood. What is new and has sparked the debate is that since the development of pharmacological antidepressants, there are *medical technologies* that can enhance mood and allegedly alter character traits. On the one hand, these medicines appear to be more acceptable to the public than many illicit mood enhancing drugs (like cocaine

or ecstasy), perhaps due to the fact they are considered to be safer and have fewer side effects than such drugs. On the other hand, the use of pharmaceuticals for the improvement of the mood and character traits in more-or-less healthy people raises both popular and academic concerns about the proper limits of medicine and about the ethics of using pharmaceutical means to alter oneself. It is nevertheless remarkable that the debate on mood enhancement has focused exclusively on pharmaceuticals and that the use of mood enhancing drugs or substances like alcohol has never been part of the discussion.

A Brief History of the Debate on Mood Enhancement

Ever since the development of psychopharmacology, there have been ethical concerns and social discussions about the use of pharmaceuticals to help people cope with stress, boredom, and anxieties. Well-known examples include the tranquilizer Miltown in the 1950s or “mother’s little helper” Valium in the 1970s (Elliott and Chambers 2004).

In the mood-enhancement debate as it has developed in bioethics and neuroethics since the 1990s, it was the new wonder drug Prozac that became the paradigmatic mood enhancer. “Prozac” has come to stand for a whole group of similar antidepressant drugs and is used as a kind of shorthand.

While the first pharmacological antidepressants were developed in the 1950s and 1960s, *Prozac* was the first of a whole new class of drugs: the selective serotonin reuptake inhibitors (SSRIs). The SSRIs were developed in the 1970s and had far less side effects than previous antidepressants. Prozac was first marketed in 1987; it instantly became a huge success and was welcomed as a kind of wonder drug. In 1994, only 7 years after its release, Prozac was the second best-selling drug worldwide.

The popularity of Prozac and other SSRIs as “lifestyle” drugs was probably increased by the 1993 book *Listening to Prozac*, in which psychiatrist Peter Kramer described how some of his patients used Prozac and came to feel *better than well*. They did not just recover from depression, but some of them also felt like certain character traits had been improved. They reported, for example, feeling less shy, more outgoing, more active, more enthusiastic, and more in control. Some even reported finally feeling like themselves. “This is who I am. I just feel strong. I feel resilient. I feel confident” (Kramer 1993, p. 219). Kramer coined the term “cosmetic psychopharmacology” to describe these effects. His book not only led to an increase in the popularity of Prozac but also started off the ethical debate.

Throughout the 1990s and 2000s, the sales of SSRI antidepressants kept rising. From 1999 to 2000, the sales of antidepressants in the United States saw a 20.9 % increase, to a figure of $10.4 billion, maintaining their position as the best-selling category of drugs (Elliott and Chambers 2004, p. 5). From 2000 onward, however, more critical reports on these drugs started to appear, questioning their alleged safety and lack of side effects. It gradually became clear that there were side effects, especially the increased suicide risk in children and adolescents using them was hotly debated in the scientific literature. Also reports of Prozac causing outbursts of

aggression and violence in rare cases started to appear, and the use of SSRIs was even presented by the defense as a mitigating factor in violence and murder trials (Healy et al. 2006). Moreover, the actual effectiveness of these drugs for the treatment of depression was increasingly questioned. A number of meta-analyses, especially the ones taking also data from unpublished trials into account, suggested that the effect of SSRIs on depression was far less than it had been believed to be (Kirsch 2010). According to some influential critics, SSRIs have only little more than a placebo effect on most people; they are not effective in mild or moderate depression and only have limited effects on people with serious depression. The type of enhancement effects on character traits and mood of more-or-less healthy people that Kramer described has never been shown to exist in a research setting – although insufficient research has been done here to make definitive claims (de Jongh et al. 2008; Repantis et al. 2009).

From being a wonder drug and the paradigmatic "mood enhancer" that sparked the ethical debate, Prozac has now become a highly contested medicine. Not much is left of the idea that pharmaceuticals can enhance mood and personality traits of healthy people as they desire. As known from the field of sociology of expectations, the high hopes and the hype that is often created around new technologies can in turn elicit a response of resistance, fear, and rejection. This dynamic can create a very polarized debate between proponents and opponents of a new technology or application that is not based on actual facts and knowledge. This seems to have happened to some degree in the mood-enhancement debate, where some of the enhancements that have been discussed – e.g., the adjustment of character traits according to one's desires or the delivery of instant happiness – have turned out to be of a rather hypothetical nature. This does not necessarily take away from the validity and importance of some of the philosophical and ethical arguments that have been made, though.

Since the mid 2000s, the ethical debate on pharmacological enhancement has turned in large part away from mood enhancement and towards cognitive and lately also moral enhancement. Some new emerging neuromodulation technologies, such as Deep Brain Stimulation (DBS), Transcranial Magnetic Stimulation (TMS), Vagal Nerve Stimulation (VNS), or Direct Current Transcranial Stimulation (DCTs), can trigger the same type of questions, however. These technologies are being tested in the area of depression, and they seem to have a potential to also enhance mood and perhaps certain character traits in healthy people. Especially DBS, with its potential to elicit manic and hypomanic states in patients, may be seen as a new type of "mood enhancer." The advent of these technologies may revive the mood-enhancement debate in the near future. At this point in time, it is difficult to predict how exactly these new neuromodulating technologies will develop. The history of SSRIs, which started out with high expectations and a hype but turned out to be at best a set of reasonably effective medicines for severe depression, and at worst "*The emperor's new drugs*" (Kirsch 2010), poses a warning against exaggerated expectations.

Be that as it may, discussions on how we should conceptualize mental problems, what the place for pharmaceuticals in dealing with such problems ought to be, and

how we should understand ourselves in relation to both mental problems and the proposed medical solutions for these problems remain important and topical.

Philosophical and Ethical Issues in the Mood-Enhancement Debate

A number of ethical and philosophical issues have come up in the debate on mood enhancement. These include worries about safety and side effects of psychopharmaceuticals, issues of justice, and fairness, who will have access, who will pay; but since these are not very specific for mood enhancement, they will not be dealt with here. The most important issues in the debate, and the most specific for the enhancement of mood and character traits, are those about the nature and value of happiness and human well-being; authenticity, identity, and the self; and the proper role of medicine and the medicalization of psychological states and traits, including the worry that the use of pharmaceuticals will divert attention from underlying personal or societal problems.

The Nature of Happiness

Is happiness initiated by a pill true happiness? Is it not somehow artificial and fake, of lesser value than real happiness? But what is real happiness and what contributes to true well-being? This is one of the primary concerns that mood enhancement by means of pharmaceuticals – or other neurotechnologies – raises.

Perhaps the most appealing illustration of this concern is the use of Soma in Aldous Huxley's Brave New World. In Brave New World everybody is engineered to do exactly what they are supposed to do and everybody is happy and content. Illness and misfortune have been banned, and people's lives revolve around pleasure and consumption. The drug Soma is widely used to help people feel good and to keep them content in this perfect world. "And if ever, by some unlucky chance, anything unpleasant should somehow happen, why, there's always soma to give you a holiday from the facts. And there's always soma to calm your anger, to reconcile you to your enemies, to make you patient and long-suffering" (Huxley 1994/1932, p. 217).

Given the dystopian features of Brave New World, it is no surprise that opponents of mood enhancement frequently invoke this image as a deterring example. The President's Council on Bioethics, for instance, uses it to illustrate "the debased value of a spurious, drug induced contentment. Soma – like cocaine, only without side effects or addiction – completely severs feeling from living, inner sensation from all external relations, the feeling of happiness from leading a good life" (President's Council on Bioethics 2003, p. 294).

Robert Nozick's well-known thought experiment of the Experience Machine has been invoked to discuss this issue and to explain why mere contentment or happy feelings may not be what we should be striving for with mood enhancement. The experiment shows that we want *reasons* to be happy and not just *causes* of happiness. We want our happiness to relate to things that actually happened, things we actually have done, or that are out there in the world and not just be caused by some pill or some machine (Schermer 2011; Kahane 2011; Liao and Roache 2011). This boils down to an argument against a hedonistic view of well-being; it shows that a happy mood is not enough for, or identical to, overall well-being, to having a good life, or to human flourishing.

Another way to make this point is to show that certain kinds of suffering may actually be essential parts of our moral lives and even of what it means to live a good life. Suffering from grief at the loss of a loved one, for example, is not the kind of suffering or the kind of negative emotion that one should want to "get rid of" by means of a mood enhancer (President's Council on Bioethics 2003; Olsen 2006). This implies that the relationship between so-called mood enhancement and well-being is much more complex than the simple idea that brightening one's mood will make you better off.

The Value of Alienation and Discontent

Another related point of critique on mood enhancement as a simple erasure of unhappy feelings and alleviation of discontent has been eloquently voiced by Carl Elliott (1998, 2003, 2004). According to Elliott, the problem that Prozac or other mood enhancements address is not so much depression or anxiety but spiritual emptiness and alienation. This alienation – which can be of a personal, cultural, or existential nature – is a symptom of Western (especially American) culture and points at deeper problems with giving meaning to life, finding a purpose, and making sense of the world we inhabit.

Using Prozac or other means to alleviate feelings of alienation does not address these deeper issues; it does not answer metaphysical questions of meaning, of who we are, or who we should be. According to Elliott, Prozac is not the answer because it wrongly defines the problem as a psychiatric one instead of an existential issue of meaning and alienation. "Imagine an accountant living in Downers Grove, Illinois who comes to himself one day and says, 'Jesus Christ, is this it? A Snapper lawn mower and a house in the suburbs? Is this all life has to offer to me?'" (Elliott 1998, p. 180). Should this man take Prozac to relieve this predicament and his feelings of alienation and being discontent with his life, Elliott asks. Should a doctor prescribe it for these reasons, to help people relieve their feelings of dissatisfaction, alienation,or meaninglesness? Elliott's suggestion is that there may be other and better ways to deal with such existential questions.

A somewhat related point of critique is that of social quietism: whereas discontent, unhappiness, or alienation could be reasons to try and change the social or cultural circumstances that cause them, taking pills to improve mood will only advert attention from those societal issues. This is also depicted by the role of Soma in Brave New World: it is distributed to the masses like bread and circuses to keep them quiet and content.

Authenticity and Identity

The issue that has raised the most extensive and fierce philosophical debate in the context of mood enhancement is that of authenticity and personal identity. Ever since Peter Kramer suggested that Prozac might not only enhance mood but also affect character traits in a favorable way, the question of what this does to the personal identity and the authenticity of the person using such drugs has been on the agenda. One of the most remarkable findings of Kramer was that his patients reported feeling actually *more themselves* when they used Prozac than without it. Can a drug bring you closer to "who you really are"? Can it enhance authenticity? Or do pharmaceuticals and other technological means that change personality traits in an artificial way make one less truly oneself, or a less authentic person?

Although proponents and opponents in this debate seem to agree that authenticity is an important value, they do use different accounts of what authenticity is (Parens 2005; Bolt 2007; Levy 2011). Some seem to refer to a notion of a true core self that lies hidden somewhere to be discovered; it refers to a way of life that is uniquely and truly our own. Others consider an authentic personality something that one has created and shaped oneself, that is not static but evolves over time. The difference boils down to seeing authenticity as either a matter of self-discovery or of self-creation.

Carl Elliot is one of the most important authors who have claimed that psychopharmaceuticals can threaten a person's authenticity. According to him, the use of such drugs may make you lose touch with who you really are. "It would be worrying if Prozac altered my personality, even if it gave me a better personality, simply because it isn't *my* personality. This kind of personality change seems to defy an ethics of authenticity" (Elliott 1998, p. 182). Elliott also points to the social pressures and the drive to conform to social expectations and norms and that is behind the use of these kinds of drugs. These drugs help people live the kind of life that is considered normal, appropriate, and successful by society and to adapt to the patterns and standards required for that – without questioning these norms themselves. A similar idea is expressed in this quote from Brave New World: "I don't understand [. . ..] why you don't take Soma when you have these dreadful ideas of yours. You'd forget all about them. And instead of feeling miserable you'd be jolly. *So* jolly" (Huxley 1994/1932, p. 82). The answer Bernard, the protagonist, gives to this question clearly refers to a notion of authenticity: "'I'd rather be myself,' he said. 'Myself and nasty. Not somebody else, however jolly'" (Huxely 1994/1932, p. 80). The basic idea that many people would agree with is that being yourself is a good thing even if this self is not in every respect perfect or agreeable.

The opposing view has been most forcefully argued for by David DeGrazia (2004, 2005). DeGrazia denies that there are any core characteristics that are somehow inviolable, that if changed, would render us inauthentic. According to his view, we may use psychopharmaceuticals or other means to help us change our traits to become who we want to be, as long as the choice to do so is autonomous. Pharmaceuticals can help us to shape and create our personalities, which is something we have always been doing anyway: we try to give form and direction to our lives and to who we are. Whether we use technological tools like pharmaceuticals to achieve this does not really matter. What matters is whether the changes are autonomous: whether the person in question wants them and truly identifies with them. Moreover, in order to be authentic, one should also be honest towards others about who one really is and not pretend or deceive them. “Any self-creation project that is autonomous and honest is ipse facto authentic,” he states (DeGrazia 2005, p. 112). So according to DeGrazia and others who follow this line of reasoning, if a person autonomously chooses to use mood enhancers to alter a certain aspect of himself, if people use them to bring their personality in line with who they truly wants to be, this can contribute to their authentic self-creation.

“Aspirin for the Mind” and the Mechanization of the Self

Contrary to the position of DeGrazia sketched above, some authors have argued that it *does* matter which *means* one uses to change mood or character traits. They claim that there is a significant difference between using means that are responsive to reasons, like psychotherapy, and means that work in a purely “mechanic” way, like pharmaceuticals. By using techniques that “bypass” reason or rational capacities to change traits, moods, or emotional responses, we regard persons in an overly mechanistic way, according to this line of argument (Freedman 1998; Levy 2007; Svenaeus 2009; Kahane 2011). Using pills to “cure” unfavorable emotional states makes them seem like simple headaches, without any meaning or content. However, emotional states, character traits, and behaviors are related to reasons and meaning and should not be treated with a mere “aspirin” but in a way that engages reason. What is ultimately at stake here, according to Carol Freedman, is “the perception of ourselves as responsible agents, not machines” (Freedman 1998, p. 136).

Against this point of view Neil Levy has argued that at least in certain cases, pharmaceuticals are used to change or cure emotional states or responses that, according to the person herself, have no rational ground. “We bypass her rational faculties, but we do so to treat an illness that in itself, by her own reckoning, bypasses reason” (Levy 2007, p. 117). If, however, pills would be used to create a happy mood without there being any reason for us to be happy, one could argue that the mechanization concern was valid – and here the argument touches on the previous discussion about the nature of true happiness and the Experience Machine.

Another refutation of the mechanization concern is that mood enhancers may not merely make one “feel good” without any reason but that they actually make one more

responsive to reasons. By changing our general affective orientation, they may help us to better appreciate the things that we have indeed reason to appreciate (Kahane 2011).

Medicalization

A final cluster of moral concerns about mood enhancement regards the question of whether we are talking here about the treatment of medical problems or disorders or whether it is enhancement of healthy people beyond their normal or natural selves – and if so, whether that is morally problematic. In order to answer this question, we would need to know how to distinguish between normal and pathological mental states or between mental health and mental illness. This is, however, notoriously difficult especially in the area of mental health since what we should regard as "normal" forms of mental functioning is to a significant degree determined by our cultural and historical context. Many contributors to the enhancement debate have argued that concepts of health and disease are inadequate, contested, and value laden and that it is therefore impossible to use them as the basis to distinguish between treatment and enhancement. Others have argued that this distinction is irrelevant altogether.

The boundaries between the domains of health and illness, normal and abnormal, tend to be flexible and shift over time, and this causes doubt about what we *ought* to regard as normal and abnormal, healthy, or pathological. In the mood-enhancement debate, concerns have been raised about the progressive extension of the boundaries of medicine and psychiatry and the further medicalization of emotional and social problems, which perhaps should be treated in different ways than with medication. Examples often referred to are the medicalization of shyness and the expansion of the diagnostic category of attention deficit hyperactivity disorder (ADHD).

What used to be seen as (extreme) shyness – constant anxiety in the presence of others, avoidance of social contact out of fear for others – has over the past decades come to be seen as mental illness: social phobia, social anxiety disorder, or avoidant personality disorder (Berghmans et al. 2011). SSRIs are used to treat these conditions. According to critics, this is the medicalization of a normal phenomenon, the pathologization of a social problem that should be solved in a different way, for example, by acceptance, by lowering our standards and expectations, or by training self-confidence. Likewise, the criteria for diagnosing ADHD have expanded over the years and have come to include more and more children, as well as adults. Children or adults who were previously seen as chaotic or easily distracted or overly energetic are now seen as having a disorder. Especially when these traits stand in the way of optimal performance, medication is prescribed to improve functioning. According to some this is a form of medicalization of underperformance (Schermer and Bolt 2011).

With mood disorders, a similar thing has happened. Not only have the meaning of and diagnostic criteria for "depression" changed and expanded over time, the indications for the use of antidepressants have expanded as well. Apart from their indication for major depression, SSRIs are also prescribed for anxiety disorders,

panic disorders, obsessive compulsive disorders, bulimia nervosa, and premenstrual distress. Moreover, there is also a significant amount of off-label use, i.e., the prescription of drugs for reasons that are not officially approved indications, like psychosocial problems, conflicts at work, marital problems, grief, or stress. One might call this the medicalization of unhappiness.

Authors who use the term medicalization in a pejorative sense do not always make clear why medicalization is problematic, especially not whether, and if so, why it should be considered as *morally* problematic. One reason to consider it as such is that medicalization may lead to the neglect of less harmful, cheaper, or otherwise better problem-solving approaches. Another reason concerns the safety and proportionality of risks and benefits when drugs are used for increasingly milder problems. Finally, it may lead to individualization of social problems and thus to unjust "blaming the victim" (Schermer et al. 2009). One of the causes of the increase in ADHD diagnoses may well be the increasing performance pressures on children at school and on adults in the workplace, and the increasing complexity and bustle of present day society. By labeling those people who have difficulty in keeping up the pace as "having a disorder," they are made to be the problem, instead of society. This may also place the focus of attention on solutions that affect the individual, instead of his social surroundings. For example, children who would do fine with some extra attention and smaller classrooms may now be given ADHD drugs instead.

Disease Mongering and the Role of Industry

A special role in the process of medicalization is played by the pharmaceutical industry, and much of the criticism on medicalization in the mood-enhancement debate has focused on this role (Healy 2004). Manufacturers of pharmaceuticals can exploit the blurred boundaries between what is normal and pathological and market their products as treatments for newly discovered or underrecognized disorders instead of selling them as enhancements, or so critics argue. Or as David Healy has claimed, in order to sell pharmaceuticals, one needs to sell diseases. The activities by which companies actively help to create new disease categories and "sell" sickness to create new markets are known as disease mongering. It has received ample attention and critique over the last decades, both in medical journals and public debate. The amounts of money spent on the marketing of SSRIs as well as the attempts of the pharmaceutical industries to present research outcomes in a more favorable light than may be warranted by the actual data are part of this process.

In part, the moral problems concerning disease mongering are the same as those concerning medicalization. Moreover, making a profit from medicines that are unnecessary and may cause harm is obviously morally problematic, as is manipulating research data. The ethical debate here turns not so much on the moral principles, norms, or judgments regarding such practices, as on the empirical substantiation of the allegations against the pharmaceutical industry.

Goals of Medicine

A final and related issue is whether mood enhancement belongs to the proper tasks of the doctor and whether it fits in with the goals of medicine. While treating diseases and disorders seems obviously to belong to the goals of medicine, there has been discussion about whether enhancing people's well-being or enhancing certain personality traits in the absence of pathology should also count as a goal of medicine. Would that not be drawing the boundaries of medicine too wide? While some argue that medicine should restrict itself to dealing with health-related problems, others believe that the "goals of medicine" may well include, or come to include, the enhancement of well-being (Brülde 2011; Synofzik 2009). Recent guidelines by the American Academy of Neurology suggest that some professional organizations indeed tend to accept enhancement as a part of medicine.

Anyway, the strength of the "goals of medicine" argument is limited; as Parens (1998) and others have argued, even if "the goals of medicine" could dissuade *doctors* from providing certain drugs, because they consider them enhancements rather than treatments, they provide no reason to forbid or forgo the use of such pharmaceuticals altogether. Other professionals might step in, or the distribution of such drugs could be left to a free market. A final pressing question then remains: how should such drugs be regulated, and should they be reimbursed by healthcare insurance schemes? This, however, is a topic that is by no means specific to *mood* enhancement and therefore will not be discussed further.

Concluding Remarks and Future Directions

The discussion on the ethics of mood enhancement fits into the wider debate on human enhancement, but it is also a debate on how we as human beings should understand ourselves, and how we should deal with our emotions and our own limitations, and what it means to have a good life. The discussion frequently touches upon the ways in which neuroscience and neurotechnologies are affecting our ideas about who we are and how we should understand ourselves, whether we should see mental problems as pathologies of the brain or as existential issues about meaning. Perennial philosophical questions about happiness, well-being, identity, and agency are all essential to the debate, as are more "mundane" questions about the power of institutions like medicine or the pharmaceutical industry.

While at present it appears as if the discussion on mood enhancement has subsided a bit – there have been no recent breakthroughs in either the mood-enhancement techniques or in the ethical arguments presented – this might change again in the future as new techniques like TMS or DBS become more widely available. Moreover, it seems likely that a broader range of emotions, mental dispositions, traits, and behaviors will be opened up for interventions, partly because of the developments in neurobehavioral science and partly because of the increasing importance that optimal mental functioning has in modern societies. The term mood enhancement may well become obsolete since what is at stake can be more aptly

described as “Biomedical Behavior and Emotion Regulation.” But whatever we are going to call it, the philosophical, ethical, and social questions discussed here under the flag of mood enhancement are here to stay for the coming decades.

Cross-References

▶ Ethical Implications of Brain Stimulation
▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
▶ History of Psychopharmacology: From Functional Restitution to Functional Enhancement
▶ Impact of Brain Interventions on Personal Identity
▶ Neuroenhancement
▶ Neuroethics and Identity
▶ Reflections on Neuroenhancement
▶ Research in Neuroenhancement
▶ Smart Drugs: Ethical Issues
▶ The Morality of Moral Neuroenhancement
▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Berghmans, R., ter Meulen, R., Malizia, A., & Vos, R. (2011). Scientific, ethical and social issues in mood enhancement. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities* (pp. 153–165). Oxford/Malden: Wiley-Blackwell.

Bolt, L. L. E. (2007). True to oneself? Broad and narrow ideas of authenticity in the enhancement debate. *Theoretical Medicine and Bioethics, 28*, 285–300.

Brülde, B. (2011). Is mood enhancement a legitimate goal of medicine? In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities* (pp. 218–229). Oxford/Malden: Wiley-Blackwell.

Buchanan, A. (2011). *Beyond humanity?* Oxford: Oxfrod University Press.

De Grazia, D. (2004). Prozac, enhancement, and self-creation. In C. Elliott & T. Chambers (Eds.), *Prozac as a way of life*. Chapel Hill/London: University of North Carolina Press.

de Jongh, R., Bolt, I., Schermer, M., & Olivier, B. (2008). Botox for the brain. Pharmacological enhancement of cognition, mood and personality traits. *Neuroscience and Biobehavioral Reviews, 32*(4), 760–776.

DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge: Cambridge University Press.

Elliott, C. (1998). The tyranny of happiness: Ethics and cosmetic psychopharmacology. In E. Parens (Ed.), *Enhancing human traits* (pp. 177–188). Washington, DC: Princeton University Press.

Elliott, C. (2003). *Better than well. American medicine meets the American dream*. New York: Norton.

Elliott, C. (2004). Pursued by happiness and beaten senseless: Prozac and the American dream. In C. Elliott & T. Chambers (Eds.), *Prozac as a way of life*. Chapel Hill/London: University of North Carolina Press.

Elliott, C., & Chambers, T. (Eds.). (2004). *Prozac as a way of life*. Chapel Hill/London: University of North Carolina Press.

Freedman, C. (1998). Aspirin for the mind? Some ethical worries about psychopharmacology. In E. Parens (Ed.), *Enhancing human traits* (pp. 135–150). Washington, DC: Princeton University Press.

Healy, D. (2004). Good science or good business? In C. Elliott & T. Chambers (Eds.), *Prozac as a way of life*. Chapel Hill/London: University of North Carolina Press.

Healy, D., Herxheimer, A., & Menkes, D. B. (2006). Antidepressants and violence: Problems at the interface of medicine and law. *PLoS Medicine, 9*(3), 1478–1487.

Huxley, A., (1994; originally 1932). *Brave new world*. London: HarperCollins

Juengst, E. (1998). What does enhancement mean? In E. Parens (Ed.), *Enhancing human traits* (pp. 29–47). Washington, DC: Princeton University Press.

Kahane, G. (2011). Reasons to feel, reasons to take pills. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities* (pp. 166–178). Oxford/Malden: Wiley-Blackwell.

Kirsch, I. (2010). *The emperor's new drugs. Exploding the antidepressant myth*. New York: Basic Books.

Kramer, P. (1993). *Listening to Prozac*. New York: Viking.

Levy, N. (2011). Enhancing authenticity. *Journal of Applied Philosophy, 28*, 3.

Levy, N. (2007). *Neuroethics*. Cambridge: Cambridge University Press.

Liao, S. M., & Roache, R. (2011). After prozac. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities* (pp. 243–255). Oxford/Malden: Wiley-Blackwell.

Olsen, J. M. (2006). Depression, SSRI's and the supposed obligation to suffer mentally. *Kennedy Institute of Ethics Journal, 16*(3), 283–303.

Parens, E. (Ed.). (1998). *Enhancing human traits*. Washington, DC: Princeton University Press.

Parens, E. (2005). Authenticity and ambivalence: Toward understanding the enhancement debate. *The Hastings Center Report, 35*(3), 34–41.

President's Council on Bioethics. (2003). *Beyond therapy. Biotechnology and the pursuit of happiness*. New York: Dana Press.

Repantis, D., Schlattman, P., Laisney, O., & Heuser, I. (2009). Antidepressants for neuroenhancement in healthy individuals: A systematic review. *Poiesis & Praxis, 6*, 139–147.

Savulescu, J., ter Meulen, R., & Kahane, G. (Eds.). (2011). *Enhancing human capacities*. Oxford/ Malden: Wiley-Blackwell.

Schermer, M. (2011). Health, happiness and human enhancement. Dealing with unexpected effects of deep brain stimulation. *Neuroethics*. doi:10.1007/s12152-011-9097-5.

Schermer, M., & Bolt, I. (2011). ADHD and the grey area between treatment and enhancement. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities* (pp. 179–193). Oxford/Malden: Wiley-Blackwell.

Schermer, M., Bolt, I., de Jongh, R., & Olivier, B. (2009). The future of psychopharmacological enhancements: Expectations and policy. *Neuroethics, 2*, 75–87.

Svenaeus, F. (2007). Do antidepressants affect the self? A phenomenological approach. *Medicine Health Care and Philosophy, 10*, 153–166.

Svenaeus, F. (2009). The ethics of self-change: Becoming oneself by way of antidepressants or psychotherapy? *Medicine Health Care and Philosophy, 12*, 169–178.

Synofzik, M. (2009). Ethically justified, clinically applicable criteria for physician decision-making in psychopharmacological enhancement. *Neuroethics, 2*, 89–102.

Smart Drugs: Ethical Issues

77

Alena Buyx

Contents

Abstract

This chapter provides an overview of the debate regarding the use of so-called smart drugs – drugs originally developed for medical purposes, which are used to improve "normal" human cognitive function. Following a brief summary of the currently available empirical data, it focuses on three areas that have been and continue to be most controversial: questions of definition from philosophy of medicine, ethical issues for individuals, and ethical aspects of smart drug use for society.

A. Buyx
Centre for Advanced Studies in Bioethics, University Hospital Münster and University of Münster, Emmy Noether Research Group Bioethics and Political Philosophy, Münster, Germany

School of Public Policy, University College London, London, UK
e-mail: alena.buyx@ukmuenster.de; buyxale@ukmuenster.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_91,

Introduction

"Smart drugs" refer to pharmaceuticals that are taken not to treat a particular illness, but to "enhance" cognitive function beyond what is usually considered normal in humans. As such, they are one example of what has come to be called (medical) enhancement, that is, the application of medical technologies and interventions to improve or optimize human capacities and traits. Medical enhancement is nothing new; examples of the nontherapeutic – for example, spiritual – optimizing application of medical technology can be found as far back as the Stone Age. The unprecedented increase in medical options available to affect human characteristics and abilities over recent decades offers more options to "perfect" humans, but the activity as such and the desires behind it are far from novel.

Smart drugs are examples of what has summarily often been called "cosmetic psychopharmacology" (a term coined by Peter Kramer, e.g., Kramer 1993). The term "smart drugs" is to some degree problematic, because it can be seen to carry the connotation that the drugs in question are themselves "smart" or that it is indeed "smart" to take them. However, with this caveat, it will nonetheless be used in this article as a neutral term and alongside the more technical phrase "pharmaceutical cognitive enhancement." The expression "smart drugs" helps to highlight that this type of enhancement does not include other medical, nondrug technologies to improve cognition, e.g., training and nutrition to "boost brainpower," or the application of transcranial magnetic stimulation (TMS), deep brain stimulation (DBS), or brain-computer interfaces (BCI) to that end (but see "Cross-References"). It also denotes that the improvement that is aimed for applies to cognitive function and not personality, mood, or morality, which are other forms of cosmetic psychopharmacology (see "Cross-References").

The ethical implications of the use of smart drugs have not only fast become one of the biggest issues in academic neuroethics and indeed bioethics; they are also a favorite topic in the popular media. Recently, there has been criticism of "hype" around smart drugs. Studies show that evidence of the prevalence of use as well as the positive effects of smart drugs is routinely overstated in news articles, while the relative paucity of data on benefits, as well as potential side effects, are often underplayed (Partridge et al. 2011; Forlini and Racine 2009). The state of evidence on effects and prevalence is also ignored in some of the academic bioethics literature (Boot et al. 2011; Quednow 2010). This can, in turn, lead to what has been termed "speculative ethics" in other contexts (Nordmann 2007), where arguments are being developed based on unrealistic assumptions or expectations and conclusions drawn that have little bearing on reality.

It is one of the goals of bioethics to stay abreast of current trends and developments in both science and society, so that potential ethical problems can be anticipated or are at least not overlooked or ignored. Partly with this in mind, an attempt will be made in the following to present a balanced account of the ethics of pharmacological cognitive enhancement. As part of this effort, however, it is necessary to give at least a brief and general overview of the empirical background; owing to the scarcity of space, this should by no means be taken as exhaustive.

The Empirical Background

What Are Smart Drugs?

From caffeine to cocaine, there are very many pharmacological agents that can have the effect of altering and sometimes improving cognitive function, several of which have been around for a long time (for overviews, see, e.g., Nutt et al. 2007; AMS 2008). This contribution focuses on a group of drugs that share a number of characteristics: They have been developed for medical use, at least initially, and have therefore, at least to some degree, been studied by the usual means of medical research, for example, randomized controlled trials (RCTs); they are in current use as prescription drugs to treat neurological, psychological, or psychiatric conditions and are as such and in these applications legal in most countries; and they are often taken to be safer and have fewer side effects than common recreational, illegal drugs (such as cocaine). These are the drugs that debates about smart drugs or pharmacological cognitive enhancers mostly refer to. The best-known ones are modafinil, methylphenidate, and some amphetamines such as Adderall (see Table 77.1).

Frequency of Use

These drugs can be used off-label (not as part of a therapy of one of the described conditions) by healthy people with the aim to improve attention, wakefulness, alertness, concentration, working and long-term memory, cognitive control, spatial planning, and other cognitive functions – that is, they might be used as cognitive enhancers when preparing for and undertaking exams and tests, writing papers, and managing tiredness, stressful and long work days, high-pressure professional situations, etc. Media reports as well as academic articles often paint a picture of this being a "widespread" or "common" practice, particularly in some populations, such as college students, and some professional groups such as pilots, academics, and scientists, with use reported in the ballpark of 20–30 % in some groups (Chatterjee 2006; Sahakian and Morein-Zamir 2007; Anon 2009; Roxby 2011, meta-reviews Ragan et al. 2013; Smith and Farah 2011). Due to the perceived urgency of an "enhancement epidemic," this in turn has lead to an intense debate around the ethical and social implications of this use for individuals and society and calls for policy making and regulation.

Recently, there have been systematic studies evaluating the available evidence on the prevalence of use of pharmacological cognitive enhancers (the newest and most exhaustive ones are, e.g., Ragan et al. 2013; Smith and Farah 2011). These report that there is still a lack of reliable, representative data on enhancement use and most studies have methodological weaknesses; often these are the studies that report high numbers of users. Most of the evidence available comes from the USA; there has been less activity to study the use of smart drugs

Table 77.1 Smart drugs discussed in this contribution

Modafinil	Modafinil, which is sold under the trade names of Vigil or Provigil, is a stimulant that is approved in many countries as a treatment for narcolepsy. It is also used to treat shift work sleep disorder and excessive daytime sleepiness associated with obstructive sleep apnea. The exact work mechanism is still unclear
Methylphenidate	Methylphenidate, best known under its trade names Ritalin or Concerta, also a stimulant, is an established treatment of attention deficit-hyperactivity disorder (ADHD) approved in many countries. It is also used in some countries to treat narcolepsy and as an additive or augmentative treatment for depression. The main work mechanism can be described as inhibiting the reuptake of the neurotransmitters dopamine and noradrenaline/norepinephrine (Heal and Pierce 2006), leading to higher levels and utility of these in the brain
Amphetamines	The best known of these is the stimulant Adderall, a composition of several amphetamine salts, which is used in the treatment of ADHD and narcolepsy. It is approved for these indications in the USA, but currently not in Europe. While the exact work mechanism differs from methylphenidate, it also acts as a reuptake inhibitor of dopamine and norepinephrine/noradrenaline

in other parts of the world so far. What quality evidence is available does not support the diagnosis of cognitive enhancement via drugs as a widespread "epidemic" in need of urgent attention. While pharmacological cognitive enhancement appears to be a real phenomenon, for example, among students, evidence suggests that actual prevalence rates are far lower than reported in popular and some academic media and range between 5 % and 15 %, with the highest rates in college students and among some professional groups and the lowest in lifetime prevalence of use in general populations (overview in Smith and Farah 2011 and Ragan et al. 2013). Ragan et al. also highlight that rates might be lower than 5–15 %, because some of the studies, for example, conflated different types of use (e.g., use for recreational purposes) or did not distinguish between use for enhancement and use simply without prescription. European studies are rarer and very different in type, but rates seem to be even lower than in the USA; in a survey of a representative sample of the German working population, less than 1 % of the employees reported that they had used stimulants at least once in their lives against tiredness and sleepiness or for better concentration (DAK 2009). Another large German study among pupils and students showed lifetime prevalence between 0.8 % and 1.6 % (Franke et al. 2010). In a UK study, 0.5 % of staff and students of a University reported having used stimulants "to study" (Holloway and Bennett 2012), while a Belgian study reported that 4 % of a large student sample reported stimulant use during exam periods (Rosiers et al. 2010, cited in Ragan et al. 2013). In sum, while there is evidently some prevalence of pharmaceutical cognitive enhancement particularly in younger populations still in education, this is not a common practice.

Evidence of Effectiveness and Side Effects

All three drugs that are the focus of this contribution have been studied with established medical research methods, including double-blind RCTs. Results of enhancement effects drawn from these studies, as evaluated in recent systematic meta-analyses (Repantis et al. 2010; Smith and Farah 2011; Ragan et al. 2013), are, in sum, inconclusive and ambiguous and on the whole smaller than is often reported; most studies also report only on single-dose application and application in situations not resembling real life.

In an extensive meta-analysis, Repantis et al. did not find consistent evidence for a cognitive enhancement effect of methylphenidate, with the exception of a positive effect on memory (especially working spatial memory). Their meta-analysis could not verify a positive effect of methylphenidate on attention. It did verify such an (moderate) effect in modafinil, but no effects on memory or motivation. There was also evidence that modafinil maintained wakefulness in sustained sleep deprivation, but did not simultaneously preserve attention and executive functions. At the same time, there was evidence of "overconfidence" and "overrating" of performance in sleep-deprived subjects. Repantis et al. question whether increased wakefulness without maintenance of cognitive function coupled with overconfidence made the application of modafinil a practical enhancement when taken with the aim to stay alert yet preserve cognitive performance at pre-sleep-deprivation levels. Generally, effects, when found, were more pronounced in sleep-deprived subjects than in healthy, rested subjects and in those with low baseline performance to begin with. This was also supported by Smith and Farah, who report some positive effects on cognitive controls in subjects with low performance and high impulsiveness, but only mixed or none for others. Smith and Farah also report positive effects of stimulants on long-term learning and memory consolidation and mixed or no results on working memory and other cognitive functions. Overall, they conclude that the enhancing effects of the stimulants in question appear to be small. Ragan et al. report similar findings, adding that personal accounts from media and online stories about real-world application often describe bigger results than those found in academic studies. This might be due to a placebo effect or to individual differences.

Side effects of smart drugs are not often reported systematically (Ragan et al. 2013). When available, reports of side effects from studies also often translate poorly into the real world, where single-dose application under controlled circumstances is rare. In most evaluated studies, the drugs in question were well tolerated, with no severe side effects. Typical side effects included headaches, dizziness, gastrointestinal complaints, tachycardia, palpitations, nervousness, sleep disturbances, insomnia, and anxiety, but these led only infrequently to dropouts (Repantis et al. 2010). Most authors agree that the stimulants are not virtually side effect-free, as is routinely reported, but do have moderate side effects. However, severe side effects appear to be rare. An exception of this could be negative psychological and personality effects, for example, on the disposition to addiction or behavior towards other people, as reported anecdotally (Guardian student blog 2012), but this is so far unclear.

In view of these empirical results, some authors have questioned whether we are actually facing a problem of smart drug use and its consequences that is significant enough to warrant the current large scale debate (Quednow 2010). However, foregoing discussion could preclude informed argument of a real phenomenon that might, after all, be emerging and thus change and develop in the future. At this stage, it appears that the most sensible approach towards the issue of smart drugs is to continue efforts to get better data and evidence, particularly on real-world prevalence, benefits, and effects, and continue having a sober, evidence-based discussion of the real and likely potential ethical and social implications of their use, mindful of the danger of simplistic argument, hype, and premature policy conclusions.

Ethical and Social Issues of Pharmacological Cognitive Enhancement

The following section provides a brief overview of some of the issues and arguments discussed in the debate around the ethics of smart drugs for cognitive enhancement. As mentioned, this has become a very active and large debate and it is impossible to do full justice here to all the subtleties of the many sub-discussions or to provide comprehensive referencing. The aim is instead to highlight some of the main lines of argumentation. It should also be noted that several of the issues discussed are not specific to pharmaceutical cognitive enhancement, but apply to other forms of medical enhancement as well.

There are many ways to structure an overview of the issues around smart drugs, but a common approach is to distinguish between questions that raise (1) issues of philosophy of science and medicine, (2) ethical questions stemming from actual or potential impacts of smart drug use on the individual, and (3) societal effects of smart drugs use.

Philosophy of Medicine: Treatment/Enhancement and Natural/Unnatural

Particularly in the earlier stages of debates around medical enhancement, many focused on the so-called treatment-enhancement distinction, which is also relevant to the issue of smart drug use (e.g., Daniels 2000). To name just one example, there are cases of severe ADHD where Ritalin is clearly the treatment of an illness, and there are clear cases of Ritalin use for enhancement outside of any medical context, such as use by professionals to give them an edge the morning before a big presentation. However, there are also cases that fall into a grey area between the two, such as the application in a child with a possible ADHD diagnosis who was doing badly in school. Is this treatment of an illness or an unethical “quick-fix” enhancement? (Singh 2005).

To broadly summarize the arguments from the debate around the treatment-enhancement distinction, particularly in medical decision and policy making, it is

sometimes assumed – often implicitly – that a clear line can be drawn between the treatment of a medical condition and an enhancement that does not aim to cure or ameliorate disease. It is further taken that this line automatically establishes what is ethically permissible, appropriate, or even mandatory to do, either in the clinical field by doctors or more generally by policy makers. However, it has been pointed out that to draw such a line is quite impossible, owing to the fact that the underlying concepts of normality, illness, and disease necessary for the distinction are themselves disputed, unclear, and/or leave grey areas of definition (Buyx 2008). For example, many activities undertaken in preventive medicine and public health are not aimed to cure an illness, but in fact to improve or optimize bodily functions beyond "normal" (i.e., heighten the body's ability of immune response via vaccinations) (Juengst 1997). There is also a continuum of interventions to improve cognitive function starting from "brain training" to exercise to drinking coffee to using smart drugs to brain surgery, and just because some of these can (also) be used to treat illnesses, this does not make them ethical as such (Singh and Kelleher 2010). Similarly, it has been stressed that even if a clear line between treatment and enhancement could be drawn, it would not be possible to draw normative conclusions just from that (Daniels 2000). After all, just because an intervention is clearly defined as not treating an illness, it does not follow that it is therefore unethical or, as others claim, an ethical imperative. In order to argue one way or the other, additional arguments are necessary, for example, about the proper goals of medicine; about what constitutes desirable and undesirable cognitive and other human traits, functions, and capabilities; or about individual or societal harms and benefits of psychopharmacological enhancement (Wolpe 2002; Harris 2009; Galert et al. 2009). This latter line of thinking has been dominant lately among those writing academically on the topic; while the debates on the best definition of concepts such as health and illness continue, efforts to demarcate enhancement and treatment in a definite way have largely subsided in recent years.

Another line of argument, throwing into relief the question what the fundamental categories, terms, and concepts of science and medicine mean and what their normative upshot could be, centers on the "naturalness" or, rather, "unnaturalness" of cognitive enhancement (e.g., Daniels 2009). Some argue that "human nature" should not be tempered with; in most of the bioethical literature on pharmacological and other enhancement, this is not framed as a religious argument (although it could be), but by reference to the fact that it is ethically superior to accept what is "given" instead of striving for what is "made" (described in Parens 2005). This is mostly coupled with additional arguments, such as concern over asymmetrical relationships between generations (e.g., Habermas 2003), but at the heart of the argument lies an understanding of nature or naturalness as imbued with particular moral value. The standard response to this is that nature is not a normative concept as such and just because something is natural, it does not mean it is superior to something "made"; neither are natural states of affairs always preferable to changes to nature (Birnbacher 2006). Nature is not benevolent, and the whole enterprise of medicine could be viewed as intervening into a "natural" course of life.

While not presupposing a strong normative idea of what is natural and therefore ethical and right, others point to a natural character, particularly in children, which it would be preferable to let unfold without pharmacological cognitive enhancement (Brock 1998); this could also be framed as an argument from safety (see below).

The debates around the normative force of categories such as “illness” or “natural” are likely to continue (see, e.g., Stein 2012), but at least with regard to cognitive enhancement, a consensus seems to have been stable for some time now that for an ethical appraisal, other issues have to be taken into account.

Ethical Issues for Individuals

Safety, Efficacy, and Risk/Benefit Assessment

Like any medical intervention, enhancement poses questions of both safety and efficacy. Smart drugs have been developed for and trialed on clinical populations with diagnosed disease and there is, as pointed out, still a lack of evidence on how they work and what risks they pose in those not so diagnosed. The current evidence seems to indicate that short-term, direct physical risks are fairly minor, but that the desired positive effects might also be quite small. Due to an absence of data, it is unclear to assess to what extent smart drug use poses long-term risks, and there are some concerns that the psychological effect on individuals (e.g., with regard to addiction), currently reported mostly on an anecdotal level, could be significant (Heinz et al. 2012). In view of this uncertainty, it would be premature to denounce smart drug use as generally unsafe and risky, but it would be equally unsound to declare it to be unproblematic in terms of safety. This situation alone warrants great caution with regard to use in children; in adults, the uncertainty should be taken into account before use is recommended as, for example, a moral obligation (Harris 2009, see also Boot et al. 2011).

In terms of an assessment of the risks versus the benefits on an individual level, another controversial ethical question is whether different risk/benefit ratios are acceptable when evaluating enhancements in healthy adults as opposed to treatments of adult patients. Depending on the illness, higher risks might be deemed acceptable in patients when they are the price of treating an illness, while risks should be lower in anything available to healthy people. The opposite view could also be taken: While patients as a vulnerable group need special protection from risks that are too high, healthy and autonomous adults should be free to choose enhancement with high risks (Greely et al 2008). It is obvious that this question cannot be answered even by reference to the best evidence; it leads directly into thorny issues of individual choice and autonomy, as set out below. Again, children constitute a special demographic; most agree that in exposing them to the still uncertain risks of smart drug use, particular caution should be exercised, due to the fact that their brain is developing at a particularly high level and the need to avoid unnecessary medication (Singh and Kelleher 2010).

Autonomy and Consent

Whether a person should be free to use smart drugs, judge their benefits, and assume their risks is the question which lies at the heart of many discussions around the ethics of smart drug use. Sketched broadly, how this question is answered depends on whether the decision to take smart drugs is regarded as an autonomous choice or not. In liberal societies, there is broad agreement that as long as a person's choice is fully autonomous, it is his/hers to make, and going against his/her choice amounts to unwarranted (and sometimes illegal) paternalism. A choice is usually deemed autonomous when it is made by a person who is, at the time, fully able to make choices (i.e., he/she is in possession of his/her normal cognitive capacities), who is fully informed about risks, benefits, and alternatives, and who is under no undue pressure or coercion (see classic text by Faden and Beauchamp 1986). Under such circumstances, a person is taken to be able to give informed consent (the established tool to ensure autonomous choices within the medical sphere). There are those who argue that an adult person who has full mental capacity is the only one to decide whether he/she can reach his/her goals in life – a successful career, feeling less stressed and tired, etc. – best by using pharmacological cognitive enhancement. As long as he/she does not harm others and as long as he/she is informed about what is currently known about risks and benefits, he/she should be free to use smart drugs anytime he/she wants. Even if evidence on potential harm is still lacking, he/she should be able to judge whether he/she is willing to assume that his/her smart drug use could have unforeseen consequences. On a strong version of this view, he/she should also be free to harm himself/herself, in case future evidence shows that smart drug use does indeed lead to significant harm. On this view, any policies that ban or regulate smart drug use would be inappropriate, because they interfere with autonomous, private decision making of individuals (Bostrom 2008).

Opponents of this view usually do not deny the right to make autonomous choices, but question whether the decision to enhance cognition is truly autonomous. In view of the lack of reliable evidence and some documented misinformation in the media about the benefits and risks of smart drugs, it could be questioned whether those using smart drugs can in fact be regularly assumed to be fully informed. It is also argued that in societies with a lot of emphasis on achievement and success, people might feel pressured to enhance themselves or might be persuaded or indirectly coerced by employers or educators to use smart drugs. The reasons given for smart drug use in empirical studies often have to do with gaining an edge over competition or with being able to adapt better to competitive environments; some authors worry that students and people in professions where having cognitive abilities beyond normal range might be an advantage, such as in soldiers and pilots, could face some indirect coercion (Hyman 2011). While there is so far only anecdotal evidence that, e.g., pilots on long-haul flights or soldiers operating drones remotely are expected to use smart drugs in order to stay awake and alert for long times, peer pressure or simply the need to be able to keep up with colleagues already using smart drugs might become more common should smart drug use spread (Greely et al. 2008).

The higher prevalence of smart drug use in students is also seen as an illustration of the fact that competitive environments might exert subtle pressure to use any means available, including pharmacological “helpers,” to succeed, thus impairing fully autonomous decision making (The Danish Council on Ethics 2011; Sandel 2009).

Again, with children, the situation is different, insofar as parents are the legally responsible decision makers. Most drug use in children occurs within the medical sphere, and many have discussed rising numbers of children being prescribed Ritalin and other psychopharmaceuticals for ADHD (e.g., van den Ban et al. 2010), with some worrying that diagnoses are widening too much. Evidence shows that adolescents and young adults, on the other hand, are groups that engage in smart drug use for enhancement purposes outside of medical settings, for example, pupils in secondary schools. The question whether adolescents make autonomous decisions about this use of smart drugs is a pressing one, and there are calls to provide this demographic with special protection, even while the issues of whether adults should be free to enhance themselves by any means they want are being discussed (Singh and Kelleher 2010).

In sum, the complex discussion around autonomy and informed consent is still ongoing and reflects underlying debates between more individualistic, liberal interpretations of how individuals should be able to make choices and those views that point out systematic and structural challenges to personal free choice in modern societies. Often, whether a call for regulation of smart drugs is issued or not depends on how the view about autonomy is combined with arguments about social impacts of smart drug use, which will be discussed below.

Personality and Authenticity

As mentioned briefly above, some are concerned that using smart drugs may alter people’s personalities. Arguments along these lines can currently not be clearly proven or disproven by evidence, but in view of the fact that personality and authenticity are considered vitally important concepts for humans, the issues discussed under these headings should be duly considered (an excellent overview of the arguments with many references is Parens 2009; see also Gordijn and Buyx 2009).

Psychopharmacological enhancements, including smart drugs, face the charge that by using them, people change their personality in that they become less authentic. (The question of authenticity is closely linked to the aforementioned debate around autonomy; indeed, authenticity could be understood as an element of autonomy.) A popular version of the authenticity argument states that for a person to be authentic, he/she must be “true to himself/herself” and feel that his/her experiences and feelings are “his/her own.” Critics of cognitive enhancement fear that by altering cognitive states through drugs, people become alienated from their own feelings and experiences and separated from the world as it really is (President’s Council 2003; Elliott 2003). Users of enhancement who have used smart drugs could be seen to have taken a shortcut through technology. They may achieve a desired goal, such as better test results or work performance, but they can no longer be described as having achieved it *themselves* nor can they claim the experience as *their own*. Hence they lose awareness of who they are and what the world is truly like.

However, others have pointed out that this criticism relies on a specific understanding of authenticity (Gordijn and Buyx 2009). An alternative view would be to regard being authentic as living a life of self-fulfillment and self-creation. It has been reported that people feel alienated from their own life and experience precisely because they are not able to behave or do what they feel would fit best with the idea they have of their own personality (De Grazia 2005). On this view, cognitive enhancement could be a way to bring someone closer to the way they would like to be – better able to focus, more alert, etc. – thus making them more instead of less authentic. Cognitive enhancement could thus be regarded as a tool people utilize to achieve abilities or states which they believe are part of who they truly are and which they otherwise have difficulties to realize and experience (De Grazia 2005, for a discussion see Gordijn and Buyx 2009).

Social Aspects of Smart Drug Use

A number of arguments are brought to bear on the debate that focuses on – currently largely potential – effects of widespread and/or unregulated use of smart drugs on society. Usually, these are not harnessed by those authors who put forward a strong autonomy view, but rather by those who point to systematic or social barriers to fully informed, free decision making when it comes to pharmacological cognitive enhancement. Their more "relational" focus on the individual situated within society seems to lend itself to a greater emphasis on social aspects.

Issues of Justice

A central concern is whether unrestricted/unregulated smart drug use could lead to more inequality between individuals and more unequal societies (Greely et al. 2008, Hyman 2011). It is well-documented that traditional tools that are considered helpful to improve personal achievement and success – e.g., additional tutoring and professional training, opportunities to network, and meditation – are more available to those already privileged within societies. First of all, knowledge of the methods is usually spread widest in well-off groups such as professionals, academics, and ambitious college students. In societies that foster academic and professional competition, access to knowledge about how to gain a competitive edge and how to be better in the workplace is a valuable currency and not equally distributed across all social groups. Moreover, like other tools to improve cognitive abilities, smart drugs, if supplied on the free market, cost money. Thus there are those who worry that if the already privileged improved themselves further through the use of pharmacological cognitive enhancement, the less fortunate could fall even more behind, leading to a highly divided and unequal society, with "super achievers" at the top end and "normals" at the bottom. This has moved some to call for a ban or at least strict regulation on smart drug use and other forms of enhancement (The Danish Council of Ethics 2011). However, others have responded to such thoughts differently: Famously, some have argued that smart drugs (and other medical enhancement) could also be used to compensate for some of life's unfairness,

if those who are disadvantaged or less endowed with marketable cognitive skills were given access to them. On this view, smart drugs would be targeted specifically at and provided to those who are underachieving, for example, through a publicly funded health-care system or in schools, to help underprivileged students (Buchanan et al. 2000). Again, this is an ongoing discussion, which will most likely benefit from further research on actual social impact of smart drug use as well as further discussion on how issues of justice could be translated into public policy (see, e.g., Dubljevic 2012).

Changes to Society: Medicalization and Disengagement

Some of the issues mentioned in section "The Empirical Background" can also be framed with regard to their social effects. In societies that value success and competitiveness, with learning and work environments that encourage achievement and that are structured around high performance at particular times in order to advance (e.g., student placement tests, entry exams), some fear that smart drug use, if deregulated, could become widespread. This in turn could lead to a number of potential negative effects not only on individuals and their ability to make free and informed choices but also on societies (Danish Council of Bioethics 2011): Standards of normality could shift if significant percentages in certain groups used smart drugs. Attitudes towards conventional human abilities could change in the long term, with average abilities becoming less the norm than a negative exception. Some are concerned that people could become afraid that their normal traits are fundamentally inadequate (e.g., Sandel 2009). As a result, cognitive enhancement could trigger a medicalization of thus far ordinary human abilities.

Another effect on society which is discussed negatively is a variation of the authenticity concern mentioned above (as well as of the old adage "no pain, no gain"). If significant numbers of people enhanced their cognitive abilities and took a "shortcut" to better concentration, higher alertness, and better memory, would they still get the same satisfaction from their achievement, or would this lead to a devaluation of hard work and generate less engagement with the world? (Danish Council of Bioethics 2011, see also Parens 2009) It has been argued that widespread cognitive enhancement could lead to a society where fewer people became active in civic institutions (President's Council 2003). Others have pointed out, using analogies from professional sports, that enhancement corrupts the "rules of the game" and makes it pointless (Sandel 2009); seen in this light, smart drugs could be a means to corrupt the rules of cognitive achievement and professional success, which would be particularly troubling if some of the justice concerns materialized at the same time.

Conclusions and Directions for Future Research

On most of the ethical issues concerning the individual, discussions are still ongoing and it is quite likely that in different countries different conclusions

will be drawn, which might lead to differences in policy making with regard to access and regulation or deregulation of smart drugs. In some countries, policies around the medical prescription of the drugs in question are being clarified. However, these will be applicable mainly to the medical field of practice and will therefore not address many of the issues this contribution has focused on, which occur mostly outside of medical practice. Off-label use, somewhere between the treatment of illness and cognitive enhancement purposes, remains in a regulatory grey area, as does use that aims for enhancement from the start. Thus there is a need not only for continuing debates and research around the issues sketched in this article but also for far more consideration of the question if and how these issues could be tackled by and translated into proportionate and effective governance and policy (Bostrom and Sandberg 2009; British Medical Association 2007).

The same holds for the social effects described above. In some earlier criticism of smart drugs, potential social effects of pharmacological cognitive enhancement were expressed quite strongly without reference to their mostly speculative nature, and there was a danger of painting a picture of very negative consequences of a phenomenon that was – and to some extent, as described above, still is – in its rather early stages. However, this does of course not warrant a dismissal of the concerns around social effects, especially when regulation and policy are considered. Whether deregulation would lead to increasingly widespread use and whether that in turn could have some problematic effects on society are reasonable and important questions for both ethicists and policy makers to ask. Cognitive enhancement by other means than the drugs discussed here is not novel, nor is drug use as such. Looking at similar developments in the past and present could help point out areas of justified concern and of unwarranted overanxiety. At the same time, it would be desirable to see more specific empirical work done on potential social effects of smart drug use to fill in the existing gaps in the available evidence; certainly, the introduction of any future policies that regulate access to smart drugs would need to be accompanied by research on related individual as well as social impacts.

Cross-References

▶ Attention Deficit Hyperactivity Disorder: Improving Performance Through Brain–Computer Interface
▶ Ethical Implications of Brain Stimulation
▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
▶ Ethics of Pharmacological Mood Enhancement
▶ The Morality of Moral Neuroenhancement

References

Academy of Medical Sciences (2008). *Brain science*, Addiction and Drugs. http://www.acmedsci.ac.uk/p99puid126.html.

Anon (2009). *One in ten takes drugs to study*. Varsity 2009 http://www.varsity.co.uk/news/1307.

Boot, B. P., Partridge, B., & Hall, W. (2011). Better evidence for safety and efficacy is needed before neurologists prescribe drugs for neuroenhancement to healthy people. *Neurocase, 18*(3), 181–184.

Bostrom, N. (2008). Smart policy: Cognitive enhancement in the public interest. In L. Zonneveld, H. Dijstelbloem, & D. Ringoir (Eds.), *Reshaping the human condition: Exploring human enhancement* (pp. 29–36). The Hague: Rathenau Institute.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311–341.

British Medical Association (2007). *Boosting your brainpower: Ethical aspects of cognitive enhancements*. A discussion paper from the British Medical Association.

Brock, D. (1998). Enhancements of human function: Some distinctions for policymakers. In E. Parens (Ed.), *Enhancing human traits: Ethical and social implications* (pp. 30–48). Washington, DC: Georgetown University Press.

Buchanan, A., Brock, D. W., Daniels, N., & Wikler, D. (2000). *From chance to choice. Genetics and justice*. Cambridge: Cambridge University Press.

Buyx, A. (2008). Be careful what you wish for? Theoretical and normative aspects of wish-fulfilling medicine. *Medicine, Health Care and Philosophy, 11*(2), 133–143.

Chatterjee, A. (2006). The promise and predicament of cosmetic neurology. *Journal of Medical Ethics, 32*, 110–113.

DAK (2009). *Gesundheitsreport*. DAKForschung. http://www.dak.de/content/filesopen/Gesundheitsreport_2009.pdf.

Daniels, N. (2000). Normal functioning and the treatment-enhancement distinction. *Cambridge Quarterly of Healthcare Ethics, 9*(3), 309–322.

Daniels, N. (2009). Can anyone really be talking about ethically modifying human nature? In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (pp. 25–42). Oxford: Oxford University Press.

DeGrazia, D. (2005). Enhancement technologies and human identity. *The Journal of Medicine and Philosophy, 30*, 261–283.

Dubljevic, V. (2012). Principles of justice as the basis for public policy on psychopharmacological cognitive enhancement. *Law, Innovation and Technology, 4*(1), 67–83.

Elliott, C. (2003). *Better than well: American medicine Meets the American dream*. New York: Norton.

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. New York: Oxford University Press.

Forlini, C., & Racine, E. (2009). Disagreements with implications: Diverging discourses on the ethics of non-medical use of methylphenidate for performance enhancement. *BMC Medical Ethics, 10*, 9.

Franke, A. G., Bonertz, C., Christmann, M., Huss, M., Fellgiebel, A., Hildt, A., Lieb, K. (2010). Non-medical use of prescription stimulants and illicit use of stimulants for cognitive enhancement in pupils and students in Germany. *Pharmacopsychiatry 43*:1e7.

Galert, T. et al. (2009) *Das optimierte Gehirn*. Gehirn & Geist 11/2009, https://www.wissenschaft-online.de/sixcms/media.php/976/Gehirn_und_Geist_Memorandum.pdf.

Gordijn, B., & Buyx, A. (2009). Neural engineering – The ethical challenges ahead. In J. Giordano & B. Gordijn (Eds.), *Scientific and philosophical perspectives in neuroethics* (pp. 283–301). Cambridge: Cambridge University Press.

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456*(7223), 702–705. doi:10.1038/456702a.

Guardian Student Blog (2012). http://www.guardian.co.uk/education/mortarboard/2012/oct/24/smart-drugs-would-you-try-them

Habermas, J. (2003). *The future of human nature* (trans: W. Rehg, M. Pensky, and H. Beister). Cambridge: Polity.

Harris, J. (2009). Enhancements are a moral obligation. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (pp. 131–154). Oxford: Oxford University Press.

Heal, D. J., & Pierce, D. M. (2006). Methylphenidate and its isomers: Their role in the treatment of attention-deficit hyperactivity disorder using a transdermal delivery system. *CNS Drugs, 20*(9), 713–738.

Heinz, A., Kipke, R., Heimann, H., & Wiesing, U. (2012). Cognitive neuroenhancement: False assumptions in the ethical debate. *Journal of Medical Ethics, 38*(6), 372–375. doi:10.1136/medethics-2011-100041. Epub 2012 Jan 6.

Holloway, K., & Bennett, T. (2012). Prescription drug misuse among university staff and students: A survey of motives, nature and extent. *Drugs Education Prevention Policy, 19*, 137–144.

Hyman, S. E. (2011). Cognitive enhancement: Promises and perils. *Neuron, 69*(4), 595–598. doi:10.1016/j.neuron.2011.02.012.

Illes, J. (2005). *Neuroethics: Defining the issues in theory, practice, and policy*. Oxford/New York: Oxford University Press.

Juengst, E. T. (1997). Can enhancement be distinguished from prevention in genetic medicine? *Journal of Medicine and Philosophy, 22*(2), 125–142.

Kramer, P. D. (1993). *Listening to Prozac: A psychiatrist explores antidepressant drugs and the remaking of the self*. New York: Viking.

Maher, B. (2008). Poll results: Look who's doping. *Nature, 452*, 674–675.

Nordmann, A. (2007). If and then: A critique of speculative nanoethics. *NanoEthics, 1*, 31–46.

Nutt, D. J., King, L. A., Saulsbury, W., & Blakemore, C. (2007). Development of a rational scale to assess the harm of drugs of potential misuse. *Lancet, 369*, 1047–1053.

Parens, E. (2005). Creativity, gratitude, and the enhancement debate. In J. Illes (Ed.), *Neuroethics in the 21st century* (pp. 75–86). Oxford: Oxford University Press.

Parens, E. (2009). Toward a more fruitful debate about enhancement. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (pp. 181–198). Oxford: Oxford University Press.

Partridge, B. J., Bell, S. K., Lucke, J.C., Yeates, S., Hall, W. D. (2011). Smart drugs "as common as coffee": Media hype about neuroenhancement. *PLoS One, 6*(11), e28416. doi:10.1371/journal.pone.0028416.

President's Council on Bioethics. (2003). *Beyond therapy: Biotechnology and the pursuit of happiness*. New York: Dana Press.

Quednow, B.B. (2010). Ethics of neuroenhancement: a phantom debate. *BioSocieties, 5,* 153–156.

Ragan, C. I., Bard, I., & Singh, I. (2013). What should we do about student use of cognitive enhancers? An analysis of current evidence. *Neuropharmacology, 64*, 588–595.

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*, 187–206.

Roxby, P. (2011). *Do 'smart drugs' really make us brainier?* BBC Health News, 3. April 2011, http://www.bbc.co.uk/news/health-12922451.

Sahakian, B., & Morein-Zamir, S. (2007). Professor's little helper. *Nature, 450*(7173), 1157–1159.

Sandel, M. J. (2009). *The case against perfection*. Cambridge: Harvard University Press.

Singh, I. (2005). Will the "Real Boy" Please behave: Dosing dilemmas for parents of boys with ADHD. *The American Journal of Bioethics, 5*(3), 34–47.

Singh, I., & Kelleher, K. J. (2010). Neuroenhancement in young people: Proposal for research, policy, and clinical management. *AJOB Neuroscience, 1*(1), 3–16.

Smith, M. F., & Farah, M. J. (2011). Are prescription stimulants "smart pills"? The epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological Bulletin, 137*(5), 717–741.

Stein, D. J. (2012). Psychopharmacological enhancement: A conceptual framework. *Philosophy, Ethics and Humanities Medicine, 7*, 5.

Talbot, M. (2009). http://www.guardian.co.uk/science/2009/sep/20/neuroenhancers-us-brain-power-drugs.

The Danish Council of Ethics (2011). *Medical enhancement*. http://www.etiskraad.dk/en/Udgivelser/BookPage.aspx?bookID=%7BCBE7B949-DD1F-4696-9829-90897B2C4B71%7D

van den Ban, E., Souverein, P., Swaab, H., van Engeland, H., Heerdink, R., & Egberts, T. (2010). Trends in incidence and characteristics of children, adolescents, and adults initiating immediate- or extended-release methylphenidate or atomoxetine in the Netherlands during 2001-2006. *Journal of Child and Adolescent Psychopharmacology, 20*(1), 55–61.

Wolpe, P. R. (2002). Treatment, enhancement, and the ethics of neurotherapeutics. *Brain and Cognition, 50*(3), 387–395.

Ethics of Brain–Computer Interfaces for Enhancement Purposes

78

Fiachra O'Brolcháin and Bert Gordijn

Contents

Abstract

This chapter outlines two key ethical issues associated with the possible development of brain-computer interfaces (BCIs) for enhancement purposes. Following a brief introduction to brain-computer interfaces, a scenario in which their use for enhancement purposes becomes commonplace is sketched. General ethical issues associated with the widespread adoption of brain-computer interfaces for enhancement are then introduced. The concept of privacy is presented and various issues surrounding this concept are discussed. BCIs are likely to create new challenges in relation to informational privacy and psychological privacy. These challenges are explored, particularly in relation to liberty,

F. O'Brolcháin (✉) • B. Gordijn
Institute of Ethics, Dublin City University, Dublin, Ireland
e-mail: Fiachra.obrolchain@dcu.ie; Bert.gordijn@dcu.ie

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_89,

autonomy, personal identity, psychological well-being, and safety. It is recommended that the privacy of future BCI users is protected. Following this, the related concept of autonomy is introduced, and various issues surrounding this concept are examined. The manner in which BCIs are likely to impact autonomy is explored, with a particular focus on freedom, brain hacking, and the transfer of autonomy. Due to the moral significance of autonomy, it is recommended that restrictions are placed on the development and availability of certain types of BCIs.

Introduction

Brain-computer interfaces (BCIs) are systems that connect a person's brain with a computer and an external device. A BCI detects electrical signals in a person's brain via electrodes attached to the scalp, inserted onto the cortical surface, or via intracortical electrodes. The computer component then interprets these signals, translates them into commands, and sends these commands to the external device. In effect, BCIs enable users to control external devices with their thoughts. BCIs have developed quickly over the past few decades (Wolpaw et al. 2000). The World Technology Evaluation Center's (WTEC) report (Berger et al. 2007) on BCI development provides a comprehensive overview of research in the field. Initially, BCIs were designed for therapeutic purposes in order to help people with severe physical impairments (e.g., paraplegics or people with locked-in syndrome) to interact with the world as well as to improve the functionality of artificial limbs. Lately they have made their way into the gaming industry to enhance the interactivity of games-entertainment systems, as well as being of interest to the automotive and robotics industries (Berger et al. 2007). Their use for gaming, cars, and robotics illustrates the potential for enhancement. After all, with BCIs users can move things around exclusively with the power of their thoughts without moving a single muscle, thus extending normal motor abilities. In the chapter at hand, enhancement is understood in a very broad sense as the extension of human capabilities and skills (including capabilities to play and to imagine) that are already within the normal range beyond the standard scope. The current development of BCIs for enhancement purposes demonstrated that this technology has the potential to be used beyond the therapeutic realm. A change from BCIs only being used by a very small percentage of the population – severely impaired patients – to a state in which they are commonplace or even ubiquitous (much like the rapid adoption of personal computers and mobile phones) and are being used for enhancement purposes would present new ethical issues. After a brief sketch of some of these issues, the focus turns to two ethical issues associated with BCIs for enhancement purposes: privacy and autonomy. Both privacy and autonomy are likely to be significantly and directly affected by BCIs. An examination of these issues seems timely. It would seem only prudent to have regulations in place prior to the widespread adoption of BCIs. First though, a potential development of BCIs into a future technology with a widespread use

for enhancement purposes will be sketched. This will provide the basis for an analysis of the ethical issues associated with this scenario.

The Scenario: Widespread Use of BCIs for Enhancement Purposes

BCIs were originally developed for therapeutic ends. By translating electrical brain activity into commands that can be interpreted by a computer, BCIs have been successful in allowing severely paralyzed people to control computer cursors and external devices connected to the computer (Pires et al. 2012; Lopes et al. 2013), including robotic arms (Minati et al. 2012).

More recently, BCIs have been designed for enhancement purposes as well. They have, for example, been developed for entertainment purposes (Nijholt et al. 2009), and a number of these BCIs have already entered the market, such as the Emotiv EPOC, the NeuroSky, and the intendiX SOCI. BCIs are also likely to have military applications, with the US Department of Defense previously having shown some interest in developing the technology (McGee 2007). In addition car manufacturers have begun to research the technology to aid drivers in controlling their vehicles (EPFL 2011). There is also the possibility that BCIs will be used to interact with "Smarthomes" – homes in which information technologies are integrated with the appliances and systems so as to enable communication between them. In such homes the systems, appliances, and environment will be able to respond to electronic commands (Edlinger et al. 2011). Neurofeedback uses electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) to provide information about a user's brain activity. Neurofeedback is used to "train" the brain, in order to help it function better, for instance, to encourage relaxation (Plass-Oude Bos et al. 2010). Neurofeedback has also been used in therapeutic contexts.

As BCI technology develops and becomes cheaper and more sophisticated, more and more enhancement applications are likely to be found and their commercial potential exploited. If they continue to develop at their current pace, they might very well be a prevalent feature of society in the future. Enhancement BCIs could be extremely useful in day-to-day living, as your thoughts could control your vehicle or the devices in your home. The evolution of augmented-reality headsets (such as Google Glasses)[1] might also be combined with BCIs so that the Internet display in the headsets could be utilized via thought rather than voice command. BCIs are already being used for entertainment purposes, as mentioned above; so it seems possible that they will begin to merge with virtual reality technologies in order to provide the brain with experiences that will seem real.[2] For instance, a virtual reality headset that can "read" brain activity would be able to

[1]Augmented reality refers to scenarios in which the real, physical world is overlaid with computer-generated imagery or information, accessible via a device such as a mobile phone or Google Glasses.

[2]Indeed, virtual reality theorists have argued that the brain does not easily distinguish between the real and the virtual on a perceptual level (Blascovich and Bailenson 2011).

respond not only to the user's commands but to their subconscious thoughts, adapting the virtual world to enhance the user's experience and enabling the user to interact with the world without the necessity of having to use a mouse or keyboard. Significantly, the BCI would also have a record of the brain activity of the user. Besides enabling users to directly connect to the Internet, BCIs might perhaps even facilitate communication directly from brain to brain (cf. Gordijn and Buyx 2010).

The widespread adoption of BCIs for enhancement purposes is plausible. The phenomenal development of computer technologies means that BCIs may well become cheap and easy to use. Competition between extremely wealthy technology firms suggests that the economic conditions for their development exist, and market demand for personalized computer devices suggests the public has an appetite for novel technologies.

Ethical Issues Regarding BCIs for Enhancement Purposes

There are broad ethical concerns surrounding the harms and benefits of BCIs and their impact on society. They also have interesting ethical implications in relation to responsibility, privacy, and autonomy. This chapter will focus on privacy and autonomy and only briefly mention the other concerns. Regarding the harms and benefits of BCIs and their impact on society, it is difficult to avoid speculative claims, while the impact of BCIs on responsibility has been explored more comprehensively elsewhere (O'Brolcháin and Gordijn forthcoming). Firstly, some general ethical issues associated with BCIs are explored before concentrating, in the next sections, on privacy and autonomy.

Responsibility: Assuming that BCIs become more generally used, there are some serious issues of responsibility to be resolved. It is unknown, thus far, how easily people will be able to control external things via their thoughts. There is a possibility that training and licenses will be required. Also, the allocation of responsibility for the actions of a BCI-controlled device will be far more complicated than ascribing responsibility for normal bodily actions as the computer interpreting the brain signals and transmitting them to the device adds an extra layer in the causal chain of events determining any action the user intends. As such, the question of whether the device is "integrated into the self-concept and the structures of decision-making and acting" (Clausen 2008, p. 1497) needs to be determined. Even then, there exists the possibility of the transmission of the user's intent being garbled or misinterpreted by the computer, known as the responsibility gap, a phrase originally advanced in relation to the increased use of computers in everyday life and essential systems (Matthias 2004). This concerns the problem of allocating responsibility for the actions of a machine or device that uses genetic algorithms or some form of learning to change behavior in a way that is neither predictable nor altogether controllable. The result is that it might be difficult to ascribe responsibility to the user of the device. While Tamburrini argues that situations may arise when neither the user nor the programmers or designers of

a BCI can be held morally blameworthy for a damaging act performed by a brain-actuated mobile device (Tamburrini 2009), others (Clausen 2009) (Holm and Voo 2011) (Grübler 2011) question the significance of the responsibility gap. For instance, Grübler argues that the lack in causal control does not hinder the ascription of moral responsibility, while Clausen (2009) argues that the responsibility gap is not a practical problem in relation to BCIs as certainty regarding actions is not always necessary for ascribing responsibility.

Nonetheless, it will be imperative to determine the ways in which BCIs respond to subconscious brain activity and how responsibility will be allocated in such scenarios. The potential of hackers to utilize BCIs to either control the device or, more speculatively, the person, would seriously reduce responsibility – a person whose device has been hacked cannot be held responsible for the actions of this device; if a hacked device could be used to control a person via manipulation of their brain states, then this person could not be held responsible for "their" actions. However, if BCIs increase access to information and a person's ability to process it, then a person might be held responsible for more of their actions, given that they would have more information about the likely outcomes of said actions. Finally, BCIs might aid the allocation of responsibility (assuming the responsibility gap problem can be overcome) in that records of brain activity created by the BCI and aid investigators in determining the intentions of a user (see O'Brolcháin and Gordijn forthcoming, for a more detailed analysis).

Harms and benefits: It will be important to ensure that the use of BCIs has a positive balance of benefits and risks. Certainly, in the cases of therapeutic use of BCIs, the benefits are obvious and may appear to outweigh the risks in many scenarios. The greater autonomy offered by BCIs to those with locked-in syndrome (LIS) or who are severely paralyzed will be of enormous benefits to such people. There are ethical problems surrounding the obtaining of consent from LIS patients, as it is difficult to determine to what extent such patients are able to signal their consent or otherwise. Patients would suffer harm if the researcher misinterprets ambiguous signals in a manner favorable to their research. Moreover, raised expectations of patients regarding the capacities of BCIs might result in harmful disappointments (Haselager et al. 2009). Other harms include the risk of infection resulting from the interface, hacking, as well as threats to responsibility, autonomy, and privacy. The proportionality of the benefits and harms of BCIs-for-enhancement purposes is less clear-cut. First, there is the risk that people might find BCI use addictive. This would be particularly relevant where BCIs are used for entertainment purposes, as a BCI in conjunction with virtual reality glasses might provide experiences that are so immersive, certain people might prefer the virtual world to their own regular lives. BCIs that directly stimulated the auditory, visual, and tactile centers of the brain would be able to provide deeply immersive experiences. Using neurofeedback, such entertainment BCIs, would be able to adjust settings of the experiences they provide in response to the user's brain state. This last scenario would be more likely if BCIs were combined with other technological developments such as virtual reality goggles and/or some form of direct brain stimulation (Lecuyer et al. 2008). Then there might be long-term detrimental

neurological effects, which are currently still unknown. It would be wise to undertake long-term studies of the neurological effects of regular and repeated use of BCIs. Furthermore, the ease of access to huge amounts of information facilitated by BCIs can be a source not only of benefits but also of harms. The benefits of being able to access information from external sources (feedback from the ambient environment, Internet sources, and, possibly, drones) are obvious. However, the availability of such information might consume a great deal of our attention. Indeed, such an abundance of information might reduce our need, and hence ability, to think critically or evaluate the worth of various sources of information.

Detrimental societal impact: Broader societal changes must also be considered, although these are much harder to assess in advance. For example, BCIs may provide their users with significant advantages that those without these gadgets lack, leading to further inequalities in society. Those with BCIs might be able to access information more quickly, assess situations with greater lucidity, and interact with more of the environment than those without. Then there is the possibility that a society might become over-reliant on using BCIs, with the result that those without BCIs might find it difficult to participate in social and institutional life. For example, if ambient environments become commonplace, BCIs might be essential for interacting with them. Thus those without BCIs will risk exclusion, with consequent social and psychological risks. There may of course be more unintended consequences that cannot easily be foreseen, particularly if BCIs are used in conjunction with other enhancement technologies, virtual realities, or (should they come into existence) artificial intelligences.

However, in this chapter the focus is on two ethical issues for which BCIs almost certainly will have important implications, privacy and autonomy. While societal impacts and harms and benefit are harder to predict, it is extremely likely that privacy and autonomy (along with responsibility) will be directly affected by BCIs.

Privacy

Concept

As ICT advances it comes into greater and greater conflict with conceptions of privacy. Biometric identification uses eyes, faces, and fingerprints to identify individuals. Infrared scanners are capable of gathering information about the interior of a building; the proliferation of satellites in the skies and cameras in our cities allows populations to be watched; Internet firms such as Google and Facebook profit from gathering information about their users; mobile phones can be used to track an individual. Security fears, commercial interests, and a love of convenience have meant that contemporary society has begun to accept the erosion of privacy. On the other hand there is growing awareness of the dangers of such a development.

Widespread adoption of BCIs would add a new dimension to existing privacy threats. As mentioned, the computer in a BCI interprets a person's brainwaves and translates them into commands that are transmitted to the external device. Consequently, the computer will create a record of these signals. In the future, as BCIs become more sophisticated, they might possibly enable a record of people's intentions, emotions, and thoughts. Currently of course, a record of thoughts would be impossible – BCIs presently only enable limited actions, such as the movement of a cursor on a computer screen or the sending of steering directions to a car. To repeat, however, as BCIs grow in complexity and as the relation between brain activity and thought is better understood, BCIs may enable the creation of a record of the user's thoughts, emotions, and intentions. The computer component of the BCI is required to translate brain signals to the device, whatever it may be. As a result, a record of that brain activity will be kept. Depending on the accuracy with which specific brain signals can be interpreted as thoughts, emotions, or intentions, the computer will be able to keep a record of these states. This might radically reduce the degree of privacy an individual possesses. However, it is first necessary to present an outline of the concept of privacy.

Privacy plays a role in the development of liberal thought, with both Locke (1689) and Mill (1859) defending the rights of people to a sphere from which the government was excluded. Locke, in his defense of individual rights, argues that governments are not permitted to interfere with a person's property, including the property in their own person (Locke 1689). Locke's insistence that a person's property must be free from interference without consent implies a respect for privacy. Certainly Locke would argue that a government has very limited rights to interfere with a person's body without their consent. While Locke focuses on property, Mill argues that liberty is of the utmost importance, as it allows people to develop their own capacities. As such, it is essential that governments and others are not allowed to interfere with the individual so long as the individual is not harming others (Mill 1859). Both Locke and Mill emphasize the importance, within a liberal society, of a space in which a person is free from interference, in short, a private space. While privacy is not identical with liberty, it is intimately connected with it.

In contemporary discussions it is important to distinguish different kinds of privacy. For instance, a distinction can be made between informational privacy (confidentiality, anonymity, secrecy), physical privacy (modesty and bodily integrity), and associational privacy (relating to sharing of intimate moments, i.e., during illness) (Allen 2011). It has also been suggested that there ought to be a category for psychological or mental privacy (Floridi 1999). This would concern freedom from psychological interference or intrusion.[3] Of course, it was initially supposed that

[3]The categories of privacy mentioned above plausibly overlap, particularly where BCIs are concerned. For instance, someone using a BCI to "hack" a mind would obviously breach informational privacy by gaining information to which they have no right. This information might also concern intimate relations of that person with another, thus, breaching associational privacy. Finally, from a materialist perspective, hacking the mind, removing some information from the brain is breaching a person's physical privacy.

psychological privacy was guaranteed in part because of others' inability to access the contents of the mind. With BCIs, of course, this assumption no longer holds. Our concern is with informational and psychological privacy, as these are the types of privacy most obviously affected by BCIs. Furthermore, while the majority of the above conceptions of privacy might be culturally specific and may not be valued in smaller-scale societies and may not always have been present historically, the privacy of thoughts, intentions, and emotions has thus far been a fact of human nature. That BCIs might affect this form of privacy is worthy of serious consideration.

Issues

There are a number of ways in which the use of BCIs might conflict with informational and psychological privacy. BCIs will most likely have some record of the brain activity of the user as the computer will need to interpret brain signals in order to translate and transmit them to the device. Depending on the sophistication of analytic techniques, such data will be available to determine what people are thinking, feeling, imagining, intending, and more generally experiencing from their own subjective point of view about at various times and in various situations. Currently people can be tracked via mobile phones and satellites, our interests can be determined through analysis of our purchases, and databases and social networks store much of our histories. Despite this, our thoughts remain private, accessible only to us and beyond the reach of others. BCIs, by keeping a record of the user's brain activity (and arguably their subjective experiences), make it far more difficult for the user to control who has access to the contents of their mental life and to restrict access to those contents. With a BCI, that brain activity and the information available as a result of it is technologically available. This evolution of technology will mean that there is a greater potential for practically private realms (i.e., the mind) to be accessed. The potential loss of informational and psychological privacy has implications for autonomy, dignity, and liberty.

Both the moral status and extent of privacy can be questioned, however. For instance, Persson and Savulescu (2012) contend that the moral right to privacy is doubtful. They argue that people gaining information about you *need* not change your state; it only involves a change in the state of those gathering the information. They claim that the moral right to privacy is not established sufficiently. They do hold that one might be entitled to expect privacy in practice, due to the existence of property rights (which they take for granted). If you do something within your own property, people would only be able to find out about it by violating your property. Hypothetically, people might be able to find out about your actions within your own house without violating your property by, for instance, developing telepathy or X-ray vision. So, they argue that citizens cannot demand, as a matter of moral rights, that other people do not gain specific information about them. Furthermore, they hold that even if the knowledge that people might know something about you did involve an unpleasant change in your state (i.e., stress), this does not violate

a moral right. Moreover, they suggest that the remote probability of a grave threat to society could justify surveillance of its citizens if this can be done without violating other rights, such as rights to bodily integrity or the right to property (Persson and Savulescu 2012).

However, this stance is unconvincing. Privacy – particularly informational and psychological privacy – is important. This perspective is not uncommon. In international human rights law, privacy is generally considered to be a human right.[4] Furthermore, James Griffin (2008), for example, in his defense of human rights, argues that without privacy, autonomy is threatened and that therefore privacy should be considered a human right. Privacy is important for autonomy as the freedom from scrutiny, judgement, and pressure to conform supply the time and space in which a person can authentically choose how to live. He contends that because humans are social creatures people will self-censor (often unconsciously) if they feel that they are being watched. So if decisions about how to live were open to public scrutiny, then people's instincts for self-censorship and self-defense would come to the fore. Specifically, Griffin argues that the right to informational privacy – a provision that protects us against people's access to certain knowledge about us – can be considered a human right (Griffin 2008). Informational privacy "is a necessary condition of normative agency, and so is instrumentally valuable" (Griffin 2008, p. 236). Informational privacy covers data, conversations, and works of art if they are self-revealing (i.e., reveal the innermost thoughts and emotions of an agent) (Griffin 2008). Similarly, Thomas Scanlon contends that an account of privacy must protect a person's interests in not being seen or overheard or observed as well as the broader interests in which a person can carry out their activities without having to continually worry about other observers (Scanlon 1975). For Scanlon, the essential point is that there should be some system of limits to observation that is generally understood and observed. Privacy sets limits on the sort of information about you that others can use and the ways in which they can use it. Privacy is connected to the norms, conventions, and laws that protect us against certain kinds of offensive intrusions involving observation of our bodies, our behavior, and our interactions with other people. Moreover, Scanlon argues that privacy can be violated even if people are unaware of the violation (Scanlon 1975).

In creating new ways to challenge privacy norms, BCIs will alter the circumstances in which people can expect to enjoy privacy. The widespread use of BCIs would mean that every individual who uses a BCI is no longer certain of being free from scrutiny at the mental level. Their brain activities are potentially open to scrutiny as it may be possible to flash images before these people and record their brain responses. If this were done subconsciously (or while embedded in some form of entertainment), the user would not even notice. So companies and governments, as well as hackers, would have the potential to access information reflecting an

[4]Article 12 of the Universal Declaration of Human Rights (1948) says: "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks" ("The Universal Declaration of Human Rights," 1948).

individual's thought processes. There are experimental results that demonstrate that it is possible to gain access to a user's personal details (e.g., banking information) using a BCI connection with a 10–40 % degree of accuracy (Martinovic et al. 2012; Anthony 2012). If BCIs do become more sophisticated, it is reasonable to suppose that hackers or other external parties might also become more adept at retrieving information from a BCI. This means that it will be impossible to guarantee the privacy of thought for the BCI user. Specifically, the ways in which BCIs will challenge privacy will have implications for other areas of ethical importance, such as autonomy, liberty, personal identity, psychological well-being, and safety.

Autonomy: The importance of privacy in relation to autonomy makes a strong prima facie case in favor of privacy. As mentioned above, without certain forms of privacy, people's instinct for self-censorship and self-defense would seriously undermine their ability to choose how to live their lives authentically. Thus autonomy is undermined. The knowledge that one is being, or the fear that one might be, observed is, in most cases, likely to alter behavior. Consider how people, in phone conversations and arguably in diaries, exercise some level of self-censorship – they realize that others might become aware of their views and opinions. So even a perceived lack of privacy *will* change a person's state. If it were possible to access a person's thoughts, people would be likely to attempt to exercise control over what they think about. BCIs will make it possible for others to have access to people's brain activity. BCIs would be even more invasive than reading a diary or eavesdropping, as a person could possibly gain access to subconscious thoughts as well as conscious decisions. This potential loss of psychological privacy might affect autonomy not only in terms of how people behave but also in terms of how they subjectively experience the world – a person, if they used a BCI regularly, would no longer be able to be certain that their brain activity (and, arguably, thoughts) were unobserved. This would affect peace of mind. BCIs may also provide a more direct method of affecting autonomy, which will be discussed below.

Liberty: The information produced by BCIs could also be used by those with malicious intent to impose restrictions on liberty. For instance, widespread use of BCIs will present opportunities for unscrupulous governments. Let us assume that the adoption of BCIs is voluntary rather than obligatory.[5] Suppose conditions exist in which the majority of the population regularly use BCIs and companies store and make use of the information gathered by these BCIs. In such a scenario, governments may well lay claim to this information, at least in situations in which a risk (i.e., a terrorist threat) exists. Google's transparency report reveals the number of requests it receives from government law enforcement agencies for users' private information.[6] More sophisticated research into BCI use might

[5]However, questions might arise regarding the voluntariness of the choices to use a BCI in a world in which the social and economic systems of society are designed for use with BCIs in a similar manner to the necessity, in many parts of the world, of access to the Internet. A society in which BCIs are the norm may exert enough pressure on people to use them that the choice cannot be considered voluntary.

[6]http://www.google.com/transparencyreport/

reveal correlations between dangerous or risky behavior and specific brain activity. Governments are likely to be interested in these correlations (the Los Angeles Police Department (LAPD) is already using "big data" techniques to predict future crimes[7]). If a government were to become convinced that the correlations between certain brain states and certain dangerous acts were convincingly certain, they might successfully pass legislation to enable "precautionary" or "preemptive" safety measures, i.e., detaining people with certain brain states. In principle, such preemptive measures could expand until certain brain activity or even thoughts are made illegal. "Thought crimes" could become a reality. Such a dystopia is obviously reminiscent of Orwell's *1984*. It may be unlikely, but its possibility makes a strong case for the protection of strong privacy rights for BCI users.

Personal identity: The privacy of thoughts is particularly important as it allows people freedom to think about their options, their lives, and their hopes and fears. That people's thoughts are inaccessible to others unless they wish to communicate them is a fact of human nature (thus far). It is impossible to know how people would be and how people would perceive themselves if thoughts, emotions, and intention were fully transparent. Private thoughts play a large role in people's conceptions of themselves, in who "they are," they are central to a subjective sense of oneself. Floridi's argument that a person's personal information is an important element of their personal identity (Floridi 2005) is even more persuasive in relation to BCIs and the information stemming from the operations of a person's mind. As such, the importance of private thoughts to our subjective sense of the self is therefore of great moral worth and deserving of protection.

Psychological well-being: A person's brain activity could be assessed to determine a person's response to various stimuli (i.e., landscapes, buildings, people, speeches, advertisements, political policies). It would certainly be possible for databases to keep track of people's brain states, either in a limited or unlimited form depending on what the BCI records. If third parties are able to access the databases, then there is a clear privacy risk. Many people might not wish their opinions of and responses to certain events to be known to anybody other than themselves. The possibility of others knowing what you think about a certain subject may well cause embarrassment or shame. Others would be able to find out information about how a person responds to various stimuli and, possibly, even how a user's subconscious mind operates. Even if private information is released without malicious intent, the results may be humiliating for the person and thus have a negative effect on their psychological well-being.

Safety: Other harms might, however, arise in a scenario in which BCIs are ubiquitous. For example, an observer might gain valuable information about the person and use it for self-serving or malicious ends by hacking into their BCI and

[7](see Morozov 2013)

extracting information the user wished to keep secret. The idea that there should be limits to observation not only protects the individual from embarrassing revelations but more importantly affords the individual protection from governments or other powerful interests. In the West, security fears appear to be stronger than distrust of governments, as Persson and Savulescu's argument demonstrates; but privacy must be of immense importance for those living under authoritarian regimes. The erosion of strong privacy norms in Western democracies would certainly help authoritarian regimes legitimate further invasions of their citizens' privacy as such regimes could claim they were behaving no differently than their democratic counterparts. The West would no longer provide an alternative model against which invasions of privacy could be measured.

Recommendations

Although BCIs are already in use, they are currently not widespread. This means there is an opportunity for legislators to put in place protections of privacy in relation to BCIs. BCI users should have control over their personal information and that they ought to be able to restrict all access to this information. Robust legislation would be required to ensure that BCI users may determine who has access to their information. Therefore, it is recommended that strong rules be put in place prior to the widespread adoption of BCIs. Legislators have an opportunity to learn lessons from the rise of social networking sites and Internet behemoths such as Google where legislation often lagged behind technological development, allowing such companies to set the agenda regarding privacy.

It is plausible that providers of BCIs will wish to lay claim to much of the information produced by the devices, in ways similar to the commercial strategies of Google or Facebook. It is worth remembering that there is no presumptive right to access this information. There is no reason to suppose that simply because I decide to buy and use a BCI that anyone else has a right to the information that this BCI records about my brain activity. The information produced by BCIs is more intimate and more integral to a person's sense of self than, say, their shopping habits, the musical preferences, and their political views. The information gathered by a BCI would encompass far more of a person's personality, including their subconscious patterns of thought. Claims to such intimate information, information absolutely integral to the person, fail to respect the autonomy of the person (in Kantian terms), using their personal information as a means (to increase profits or promote security) rather than respecting the person as an end. Without strong privacy protections, such instrumental perspectives on the person would be normalized.

Users will be free to sign contracts with providers, but it is recommended that privacy protections are put in place by governments prior to the contracts being created by the legal departments of BCI providers. If privacy protections are put in place prior to the widespread adoption of BCIs, any such claims to ownership of such intimate information would be easier to resist.

Autonomy

Concept

Unlike privacy, where there seems to be a lack of positive consequences, popular use of BCIs may have some positive as well as some challenging implications for people's autonomy. The concept of autonomy is philosophically complex. It plays a central role in both Kantian ethics (the rational will was considered by Kant to be autonomous and therefore capable of laying down moral laws so that the agent has authority over his or herself) (Kant 2005) and liberal political theory (Christman 2011). For John Stuart Mill, for example, it was imperative that "human beings should be free to form opinions and to express their opinions without reserve" (Mill 1859, p. 119), and in situations in which the customs or tradition of other people are the rule of conduct rather than the individual's character, "there is wanting one of the principal ingredients of human happiness, and quite the chief ingredient of individual and social progress" (Mill 1859, p. 120). Although he may not use the term, Mill is clearly emphasizing the value of autonomy. Arguably, then, autonomy is central to our conception of the agent and of normative agency, i.e. being able to deliberate and choose without external interference, manipulation or distortion.

A working definition of the concept of autonomy will be used. Autonomy can be understood as requiring authenticity or being one's own person, freedom, and a degree of knowledge about one's circumstances and regarding one's choices. In order to be autonomous then, a person will need to be able to choose for themselves how they wish to proceed; they will need access to relevant information in order to make these choices; and they will need a certain lack of constraints so that their autonomy is not hollow.

Issues

BCIs can be used to enhance aspects of autonomy in a number of ways.

Information: Access to information relating to choices is a key aspect of autonomy. BCIs can enhance an agent's capacities in this regard. Google has developed a headset that provides users with an "augmented-reality" display (as well as a camera, microphone, and GPS). The device is voice activated and allows people to access the web hands-free while going about their day-to-day life (Rivington 2013; Google 2013). It is easy to foresee how the designers of these (or similar) devices would be eager to combine them with BCIs. These could in principle provide a person with direct web access that is activated by thought alone. The knowledge available on the web would be easily available for inspection at all times. In this configuration BCIs could also provide more information about the physical world around us. The technology could connect to drones, cameras, vehicles, or to the ambient environment (a world in which networked devices are embedded in objects with the result that networks connect users and the networked

world around them), obtaining live feeds about weather or traffic or other dynamic systems. Future evolutions of this technology might perhaps even allow webpages to be sent directly to the visual part of the brain – to be presented to the "mind's eye." BCIs (in conjunction with various sensor technologies and other programs) could also be beneficial in assessing a person's health and making them aware of their situation and the impact of their choices by using records of their brain activity to help determine levels of, for instance, stress or concentration. The BCI would be able to keep a record of the actions of the person and how their brain responded. This might facilitate the creation of a "lifelog" in which the person's brain states were correlated with their actions and habits. This record could be used to provide the user with more information about themselves with which they could make more informed choices. The potential for BCIs to provide immediate access to realms of information far beyond the normal human range illustrates ways in which they might enhance a user's autonomy.

Freedom: Secondly, BCIs might increase freedom. Certainly the therapeutic use of BCIs, such as allowing paraplegics or those with locked-in syndrome to access the Internet, to control wheelchairs, or to control parts of their environment (most likely the devices in their houses) will significantly increase their autonomy. However, these uses ought not to be considered enhancement, since they aim at restoring a person to the normal range of human functioning. There are, of course, also ways in which BCIs might be used to increase human freedom for a person already within the range of normal human functioning. At the most basic level, a BCI would enable a person to manipulate parts of the external world – smart homes, drones, or robots from far afield – while leaving their hands free. Beyond this, users may be able to communicate silently with others, participate in more immersive computer games, and perhaps share mental experiences with other users. This is an increase in positive freedom – people will have a greater capacity to do more things in more ways; thus people will have more choice in how they wish to spend their lives. This could make a person's life both more interesting and more fulfilling. However, BCIs could also threaten autonomy in significant ways, namely, through brain hacking, the voluntary transfer of autonomy and coercion.

Brain hacking: BCIs rely on a connection between the user's brain and the device. This connection enables the user to send commands to the device. The existence of this connection suggests that it might be possible for commands or signals to be sent in the other direction – for a brain to be "hacked." Given that computer systems can be hacked, there is a risk that the computer system in a BCI could be hacked. If signals can be sent to the user's brain via the BCI connection, the victim might lose autonomy. Such signals need not be direct commands, it would be enough to disrupt a person's thoughts or change their state of mind or their responses to the environment for the person's autonomy to be violated. This might be done for commercial reasons in order to increase the likelihood of a person choosing one product over another. Such methods could also be used as a form of cyberwarfare; BCI connections could be used to disrupt the thoughts of users or possibly disable users. Moreover, it might be possible to utilize a BCI user's brain activity without their being aware of it. Consequently, they would be participating

in activities (facial recognition projects, data-mining projects, etc.) without their knowledge and consent.

The possibility of controlling rats via electrodes implanted in their brains has been demonstrated (Talwar et al. 2002). BCIs do not *require* the implantation of electrodes in a person's brain, but by opening a channel to a user's brain, BCIs might (in the future) pose a risk that a person could be controlled, especially when BCIs would have become much more sophisticated, picking up detailed information from many different specified regions in the brain, thus allowing sophisticated third-party intrusions into the brain. One of the more disturbing aspects of this is that the person might be completely unaware that they have lost complete autonomy. If the victim behaved in a way that was at odds with their long-term goals and past desires, they would (if they regained autonomy) be aware that they had behaved out of character. Yet there is no reason to necessarily assume that they would regain the degree of autonomy needed to make such a judgement. If hacking were a possibility, there would always be a fear that any apparently autonomous decision a person made was not, in fact, their own. This threat attacks one of the fundamental bases of moral agency – the idea that an agent is not only one who acts but who alone has authority over the initiation of their actions (Buss 2008). Obviously, if a person's brain is hacked, they will not have authority over the initiation of their actions. BCIs, if they enable other agents to initiate the actions of a BCI user, undermine the user's autonomy and therefore undermine the grounds of moral agency. It will no longer be an axiomatic truth that a user's will is their own.

Voluntary transfer of autonomy: Not only will BCIs enable third parties (potentially) to hack into another BCI user's brain and influence their mental life in that way but BCIs might allow a person to transfer their autonomy to third parties as well. If this were the case, authority over ourselves would cease to be inalienable. A person could allow another to access their BCI and thereby gain a certain degree of control over them. That people will be able to transfer authority[8] over themselves, possibly at the level of thoughts, desires, and values, means an intrinsic feature of our moral agency disappears, i.e., the necessity of the agent having control over their internal self. Philosophically, the development of BCIs that could enable such a transfer will impact on our understanding of the metaphysics of agency; on a more practical level, the possibility of ceding authority over oneself or of having authority over oneself removed will potentially alter social relations and power structures significantly. For instance, it is possible to imagine a loss of autonomy being used in punitive scenarios. In cases where a prison sentence is not deemed an appropriate punishment (or where prisons are overcrowded), a person could be required to wear a BCI so that their brain activity could be monitored and, potentially, manipulated.

Coercion to use BCIs: Finally, there is the question of whether the choice itself to utilize BCIs for enhancement purposes will truly be autonomous. It is worth noting

[8]Transferring authority over oneself might be attractive to members of religious cults or for erotic reasons, for instance.

that the idea that BCIs will enhance autonomy presupposes an individualist understanding of autonomy that can, at the very least, be challenged. Certainly, the prevalence of computers and mobile phones in contemporary society was not a decision voted upon by the global populace at large. On the one hand, it could be argued that individuals decided, more or less autonomously, to use these new devices. On the other hand, the changing social landscape also demanded that these new devices were adopted, as otherwise people would no longer be able to do their jobs or partake in the life of a society that required the use of such devices. So, while there is a degree of choice – prepare to do research using computers or look for another career; use a mobile phone or risk social exclusion – this choice is heavily weighted in favor of the adoption of the new technology. If everyone is able to access information instantaneously via a BCI, anyone without one will be at a disadvantage. If BCIs are needed to interact with an ambient intelligent environment and these environments become the norm, BCIs will be required. It could be the case that BCIs will form an integral element of a person's health record or be needed to drive a car.

Recommendations

Given the practical risks that BCIs pose for autonomy and the significance of autonomy in moral philosophy and liberal political theory, as well as the practical importance of autonomy, it would be prudent to ensure that they cannot be used to weaken or undermine autonomy. Firstly, the use of information gathered from BCIs will have to be regulated. The knowledge of the ways in which a person responds to various stimuli would make it far easier to indoctrinate or brainwash that person, thereby reducing their autonomy. Secondly, it will be necessary to ensure that BCIs do not enable the hacking of a person's brain. This might mean that nonmedical BCIs that can be accessed *by others* and used by them to directly change a person's brain states without that person's knowledge or consent do not enter the market. This is not to imply that there should be a complete ban on BCIs that change people's brain states; such a ban would be absurd. Many technologies, not to mention everyday actions, change people's brain states. The concern with BCIs relates to harmful changes, such as undermining autonomy, invading privacy, manipulation, addiction brainwashing, and loss of responsibility. As such, changes to brain states can be conceived as existing on a spectrum ranging from trivial to harmful. Empirical research will no doubt be needed to determine where on this spectrum specific technologies lie and which ones should be prohibited. Such a ban would place restrictions (for economic reasons) on how BCIs are likely to be developed. Very strong and reliable firewall technologies will have to be developed to avoid third parties exploiting the information pathways between brains and computers for intrusion purposes. Ultimately, the importance of autonomy, and the risk posed to it by BCIs that might enable hacking or brainwashing, ought to limit the types of BCI that enter the market.

Conclusion

The use of BCIs for *enhancement* purposes poses a number of threats. These threats will be present in cases where BCIs are used for therapeutic purposes too, but they will not affect society at large as there are proportionally very few people that require or will potentially require BCIs for therapeutic purposes. If BCIs for enhancement purposes become prevalent or even necessary in order to function in society (in a way similar to access to a computer), the threats outlined thus far will need to be resolved. The widespread use of BCIs for enhancement purposes would of course have real benefits as well. They are likely to provide a great deal of entertainment for people, allowing people to experience virtually anything. People using BCIs will be able to instantaneously access digital information, connect with other users, and interact with the Internet (or its future equivalent) very directly and ubiquitously. BCIs may also enhance people's abilities to process information and evaluate their own thoughts, desires, and motivations.

It is essential to consider whether the benefits offered by BCIs used for enhancement purposes are proportional. The risks – particularly the risks to privacy and autonomy – are severe. The possible undermining of autonomy and privacy will have significant effects on the moral and political landscape. Both autonomy and privacy are important for normative agency. In light of the threats to autonomy and privacy, the *enhancement* benefits of BCIs appear relatively minor. Access to information is already almost instantaneous if you have access to a computer or smartphone. If one is lucky enough to have time to indulge in entertainment, there is no shortage. Certainly, it appears that BCIs might be useful – they might make housework easier, driving safer, and communications faster. However, these benefits do not appear to be proportional to the risks posed. Certainly, any regulators valuing a liberal society, particularly one in which autonomy and privacy are highly valued, should be cautious about permitting widespread use of BCIs for *enhancement* purposes before adequate safeguards have been developed that can effectively and safely tackle the challenges to our privacy and autonomy.

Therefore, a number of recommendations regarding the ethical issues associated with widespread BCI use for enhancement purposes are offered. In relation to privacy, it is contended that if BCIs become a prevalent tool, it is essential that they are limited in the amount of information they can gather about the user, that it is clear who has access to this information, and that protections are put in place for psychological privacy. Ideally, this information would either not be stored or, at worst, would be aggregated and anonymous. In relation to autonomy, it is argued that there should be restrictions on the development of BCI technologies that might facilitate the reduction or even loss of the user's autonomy. Users will need to be sure that they are safe from hacking-type attacks or viruses that might affect their mental states via the BCI. Finally, we contend that study of the possible long-term benefits and harms (such as addiction, societal disruption, and safety) of BCIs is required.

Cross-References

▶ Attention Deficit Hyperactivity Disorder: Improving Performance Through Brain–Computer Interface
▶ Biosecurity as a Normative Challenge
▶ Drug Addiction and Criminal Responsibility
▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethical Objections to Deep Brain Stimulation for Neuropsychiatric Disorders and Enhancement: A Critical Review
▶ Mind Reading, Lie Detection, and Privacy
▶ Neuroenhancement
▶ Neuromarketing: What Is It and Is It a Threat to Privacy?
▶ Research in Neuroenhancement
▶ The Morality of Moral Neuroenhancement

References

Allen, A. (2011). Privacy and Medicine. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Spring 2011)*. Stansford, CA: Stansford University. Retrieved from http://plato.stanford.edu/archives/spr2011/entries/privacy-medicine/.

Anthony, S. (2012). Hackers backdoor the human brain, successfully extract sensitive data. *ExtremeTech*. Retrieved March 22, 2013, from http://www.extremetech.com/extreme/134682-hackers-backdoor-the-human-brain-successfully-extract-sensitive-data

Berger, T. W., Chapin, J. K., Gerhardt, G. A., McFarland, D. J., Principe, J. C., Soussou, W. V., Taylor, D. M., & Tresco, P. A. (2007). International assessment of research and development. In *Brain-computer interfaces*. Baltimore: World Technology Evaluation Center.

Blascovich, J., & Bailenson, J. (2011). *Infinite reality: Avatars, eternal Life, new worlds, and the dawn of the virtual revolution*. New York: Harper Collins.

Buss, S. (2008). Personal autonomy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2008). Retrieved from http://plato.stanford.edu/archives/fall2008/entries/personal-autonomy/

Christman, J. (2011). Autonomy in moral and political philosophy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2011). Retrieved from http://plato.stanford.edu/archives/spr2011/entries/autonomy-moral/

Clausen, J. (2008). Moving minds: Ethical aspects of neural motor prostheses. *Biotechnology Journal, 3*(12), 1493–1501. doi:10.1002/biot.200800244.

Clausen, J. (2009). Man, machine and in between. *Nature, 457*(7233), 1080–1081. doi:10.1038/4571080a.

École Polytechnique Fédérale De Lausanne. (2011). Nissan teams up with EPFL for futurist car interfaces. *EPFL News Mediacom*. Retrieved from http://actu.epfl.ch/news/nissan-teams-up-with-epfl-for-futurist-car-interfa/

Edlinger, G., Holzner, C., & Guger, C. (2011). A hybrid brain-computer interface for smart home control. In J. A. Jacko (Ed.), *Human-computer interaction. Interaction techniques and environments* (pp. 417–426). Berlin, Heidelberg: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-21605-3_46.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology, 1*(1), 33–52. doi:10.1023/A:1010018611096.

Floridi, L. (2005). The Ontological interpretation of informational privacy. *Ethics and Information Technology, 7*(4), 185–200. doi:10.1007/s10676-006-0001-7.

Google. (2013). Google glass. Retrieved April 2, 2013, from http://www.google.com/glass/start/
Gordijn, B., & Buyx, A. M. (2010). Neural engineering. In *Scientific and philosophical perspectives in neuroethics*. Cambridge: Cambridge University Press.
Griffin, J. (2008). On Human Rights. Oxford: Oxford University Press.
Grübler, G. (2011). Beyond the responsibility gap. Discussion note on responsibility and liability in the use of brain-computer interfaces. *AI and Society, 26*(4), 377–382.
Haselager, P., Vlek, R., Hill, J., & Nijboer, F. (2009). A note on ethical aspects of BCI. *Neural Networks, 22*(9), 1352–1357. doi:10.1016/j.neunet.2009.06.046.
Holm, S., & Voo, T. C. (2011). Brain-machine interfaces and personal responsibility for action - maybe not as complicated after all. *Studies in Ethics*, *Law*, *and Technology*, *4*(3). doi:10.2202/1941-6008.1153.
Kant, I. (2005). *The moral law: Groundwork of the metaphysic of morals*. Oxford: Routledge.
Lecuyer, A., Lotte, F., Reilly, R. B., Leeb, R., Hirose, M., & Slater, M. (2008). Brain-computer interfaces, virtual reality, and videogames. *Computer, 41*(10), 66–72. doi:10.1109/MC.2008.410.
Locke, J. (1689). *The second treatise of government* (3rd ed.). Oxford: Blackwell.
Lopes, A. C., Pires, G., & Nunes, U. (2013). Assisted navigation for a brain-actuated intelligent wheelchair. *Robotics and Autonomous Systems, 61*(3), 245–258. doi:10.1016/j.robot.2012.11.002.
Martinovic, I., Davies, D., Frank, M., Perito, D., Ros, T., & Song, D. (2012). On the feasibility of side-channel attacks with brain-computer interfaces. In *Presented at the 21st USENIC Security Symposium*, Bellevie, WA.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. doi:10.1007/s10676-004-3422-1.
McGee, E. M. (2007). Should there be a law – Brain chips: Ethical and policy issues. *Thomas M. Cooley Law Review, 24*, 81.
Mill, J. S. (1859). *On liberty*. London: Penguin.
Minati, L., Nigri, A., Rosazza, C., & Bruzzone, M. G. (2012). Thoughts turned into high-level commands: Proof-of-concept study of a vision-guided robot arm driven by functional MRI (fMRI) signals. *Medical Engineering & Physics, 34*(5), 650–658. doi:10.1016/j.medengphy.2012.02.004.
Morozov, E. (2013, March 9). How Facebook could get you arrested. *The Guardian*. Retrieved from http://www.guardian.co.uk/technology/2013/mar/09/facebook-arrested-evgeny-morozov-extract
Nijholt, A., Bos, D. P.-O., & Reuderink, B. (2009). Turning shortcomings into challenges: Brain–computer interfaces for games. *Entertainment Computing, 1*(2), 85–94. doi:10.1016/j.entcom.2009.09.007.
O'Brolchain, F., & Gordijn, B. Brain-computer interfaces and user responsibility. In ten Have, H., & Gordijn, B. (Eds.). *Handbook of global bioethics*. New York: Springer (forthcoming).
Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.
Pires, G., Nunes, U., & Castelo-Branco, M. (2012). Evaluation of brain-computer interfaces in accessing computer and other devices by people with severe motor impairments. *Procedia Computer Science, 14*, 283–292. doi:10.1016/j.procs.2012.10.032.
Plass-Oude Bos, D., Reuderink, B., Van der Lar, B., Gürkök, H., Mühl, C., Poel, M., Nijholt, A., & Heylen, D. (2010). Brain-computer interfacing and games. In *Brain-computer interfaces*. London: Springer.
Rivington, J. (2013). Google glass: What you need to know. *TechRadar*. Retrieved April 2, 2013, from http://www.techradar.com/news/video/google-glass-what-you-need-to-know-1078114
Scanlon, T. (1975). Thomson on privacy. *Philosophy & Public Affairs, 4*(4), 315–322. doi:10.2307/2265076.
Talwar, S. K., Xu, S., Hawley, E. S., Weiss, S. A., Moxon, K. A., & Chapin, J. K. (2002). Behavioural neuroscience: Rat navigation guided by remote control. *Nature, 417*(6884), 37–38. doi:10.1038/417037a.

Tamburrini, G. (2009). Brain to computer communication: Ethical perspectives on interaction models. *Neuroethics, 2*(3), 137–149. doi:10.1007/s12152-009-9040-1.

The Universal Declaration of Human Rights. (1948). Retrieved April 2, 2013, from http://www.un.org/en/documents/udhr/

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., & Vaughan, T. M. (2000). Brain–computer interface technology: a review of the first international meeting. *IEEE Transactions of Rehabilitation Engineering, 8*(2), 164–173.

The Morality of Moral Neuroenhancement

79

Thomas Douglas

Contents

Abstract

This chapter reviews recent philosophical literature on the morality of moral enhancement. It first briefly outlines the main moral arguments that have been made concerning moral *status* neuroenhancements: neurointerventions that would augment the moral status of human persons. It then surveys recent debate regarding moral *desirability* neuroenhancements: neurointerventions that augment that the moral desirability of human character traits, motives, or

T. Douglas
Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Oxford, UK
e-mail: thomas.douglas@philosophy.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_92,

conduct. This debate has contested, among other claims, (i) Ingmar Persson and Julian Savulescu's contention that there is a moral imperative to pursue the development of moral desirability neuroenhancements, (ii) Thomas Douglas' claim that voluntarily undergoing moral desirability neuroenhancements would often be morally permissible, and (iii) David DeGrazia's claim that moral desirability neuroenhancements would often be morally desirable. The chapter discusses a number of concerns that have been raised regarding moral desirability neuroenhancements, including concerns that they would restrict freedom; would produce only a superficial kind of moral improvement; would rely on technologies that are liable to be misused; and would frequently misfire, resulting in moral deterioration rather than moral improvement.

In recent years, a number of philosophers have become interested in the nature and morality of moral neuroenhancement. These are usually taken to be interventions that aim to augment, and (expectably) succeed in augmenting, the morality of an individual or one or more of her traits. This chapter follows recent literature in discussing only moral enhancements of *human persons* or their traits. The term "moral enhancement" is henceforth used to refer only to enhancements of this sort.

People have traditionally sought to morally enhance themselves through means such as introspective reflection, engagement with literature, or calm moral discussion with others, and to morally enhance others through rational persuasion, the application of incentives or disincentives, or moral education. However, prompted by developments in the cognitive sciences, much recent discussion of moral enhancement has focused instead on moral neuroenhancement. Moral neuroenhancement can be distinguished from more traditional varieties of moral enhancement by the fact that it operates by altering brain states or processes directly, that is, not by (i) modifying the agent's perceptually accessible environment (as in the application of incentives) or (ii) by engaging the agent's deliberative capacities (as in introspective reflection or rational persuasion). Paradigmatic examples of moral neuroenhancement would operate via the administration of drugs or application of brain modulation techniques such as electrical or magnetic brain stimulation.

Moral neuroenhancements all involve increasing the morality of an individual or trait. Thus, two natural ways of classifying moral enhancements are according to the *metric* by which morality is measured and the *target* of the enhancement – that is, the entity whose morality is augmented. Possible metrics of moral neuroenhancement would include moral status, moral considerability, moral responsibility, moral understanding, moral virtue, moral desirability, moral rightness, and moral permissibility. Possible targets of moral neuroenhancement would include the human person or his character, motives, or conduct. There are various ways in which one could combine different metrics and targets to describe different varieties of moral neuroenhancement. However, two kinds of moral neuroenhancement have dominated recent discussion of this topic, namely, (1) neuroenhancements of the moral status of humans (henceforth, moral status neuroenhancements or MSNs) and

(2) neuroenhancements of the moral desirability of human character traits, motives, or conduct (henceforth, moral desirability neuroenhancements or MDNs).

Two main kinds of question have been asked about MSNs and MDNs: questions about their possibility and questions about their morality. For example, there have been discussions of whether MSNs or MDNs are possible or likely to become technically feasible. There have also been discussions of when, if ever, it would be morally permissible, desirable, or obligatory to (a) pursue the development of technologies that would or might enable MSN or MDN, (b) undergo a MSN or MDN oneself, or (c) impose such an intervention on others.

This chapter outlines some of the main arguments that have been advanced concerning the morality of MSNs (section "Moral Status Neuroenhancements") and MDNs (sections "Moral Desirability Neuroenhancements" to "Further Questions"). Discussions concerning the possibility of MSNs and MDNs are touched upon where relevant to arguments concerning their morality but are not comprehensively surveyed.

Moral Status Neuroenhancements

Moral status neuroenhancements aim to bring it about that a human enjoys greater moral status than she would otherwise have enjoyed and (expectably) succeed in realising this aim. There is disagreement regarding how to understand moral status, but one way of understanding it is as a metric of the strength and breadth of moral protections (such as moral rights or claims) enjoyed by a being. Understood thus, moral status neuroenhancements could operate, for example, by conferring additional moral rights on humans.

Recent discussion of moral status enhancements has taken place primarily within debates on the morality of general cognitive or mental neuroenhancements (section introduction in ▶ Chap. 75, "Neuroenhancement" and ▶ Chap. 80, "Reflections on Neuroenhancement," ▶ Chap. 69, "Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues" ▶ Chap. 76, "Ethics of Pharmacological Mood Enhancement"). It has in large part been stimulated by a speculative comment offered by Francis Fukuyama in a 2004 critique of transhumanism, the movement most committed to human neuroenhancement. Francis Fukuyama writes that:

> Underlying [the] idea of the equality of rights is that we all possess a human essence that dwarfs manifest differences in skin color, beauty, and even intelligence. This essence, and the view that individuals therefore have inherent value, is at the heart of political liberalism. But modifying that essence is the core of the transhumanist project. If we start transforming ourselves into something superior, what rights will these enhanced creatures claim, and what rights will they possess when compared to those left behind? (Fukuyama 2004, p. 42)

One of Fukuyama's worries here is that enhanced beings might *claim* more rights than are enjoyed by unenhanced humans. But another concern is that enhanced beings would *actually have* more rights than the unenhanced. The latter concern can be interpreted as a concern that human neuroenhancement could result in the creation of beings with greater moral status than ordinary persons (call such beings status-enhanced beings or SEBs). This proposition has been the object of substantial subsequent analysis, most notably by Allen Buchanan (2009, 2011, 2012).

Why might neuroenhancement result in the creation of SEBs? The usual explanation draws on an analogy. It is often thought that ordinary adult humans enjoy higher moral status than at least some of our nonhuman mammalian relatives. In addition, it is often thought that our elevated moral status, relative to other mammals, can be attributed to our possessing greater mental capacity than other mammals. (Precisely which mental capacities confer higher moral status on us is open to dispute, but candidates include our capacities for self-awareness, rationality, and moral agency.) However, if we enjoy greater mental capacity than our mammalian relatives in virtue of our greater mental capacity, one might expect that beings whose mental capacity exceeded our own to a similar degree would enjoy greater moral status than us, and it might seem that the neuroenhancement of humans could, at least in principle, produce such beings.

The creation of SEBs would, it has been argued, be morally problematic since the existence of SEBs would be bad for ordinary persons (e.g., Agar 2012a). Arguably, ordinary persons could be permissibly harmed for the sake of these SEBs in ways that they may not be permissibly harmed for the sake of one another. For example, perhaps persons could permissibly be used, without their consent, in medical experiments designed to aid SEBs. Or perhaps persons could be rightly excluded from the democratic institutions of the SEBs.

Three main responses have been made to this line of thought. First, some authors have questioned whether human neuroenhancements could produce beings of greater moral status than ordinary persons. For example, Allen Buchanan (2009, 2011) and James Wilson (2007, 2012) argue for a threshold account of moral status according to which there is some level of mental capacity above which all beings have the same moral status, and beyond which ordinary persons already lie (see also Savulescu 2009; Wikler 2009). Buchanan invokes Stephen Darwall's (2006) work to argue that the ability to engage in practices of mutual accountability is what takes one across the threshold, while Wilson argues that a being has crossed the threshold if it possesses Rawls' two moral powers – a sense of justice and a "capacity to have, revise, and rationally to pursue a conception of the good" (Rawls 2001, p. 19). In reply, a number of authors have questioned whether there is good reason to prefer this threshold account of moral status to (a) a scalar account, according to which moral status is continuously rising with some aspect of mental capacity, or (b) a multi-threshold account, according to which there are multiple thresholds of mental capacity

making out differences in moral status, including at least one threshold that lies above the level of mental capacity possessed by ordinary human persons (Agar 2012a; DeGrazia 2012a; Douglas 2011a; McMahan 2009).[1]

A second response has been to argue that, even if human neuroenhancements could produce SEBs, this would not be bad for ordinary persons. Buchanan (2009, 2011) argues that persons already possess enough moral status to be *inviolable* and that this protects them against being exploited in the ways that those concerned about the creation of SEBs fear. For example, because persons possess enough moral status to be inviolable, they could *not* permissibly be used, against their will, in medical experiments designed to aid SEBs. However, it has been argued in reply that the inviolability enjoyed by ordinary persons is not absolute, so that SEBs could be more inviolable than us (Agar 2012a; Douglas 2011a; McMahan 2009). Thus, for example, when faced with a choice between killing an ordinary person and killing an SEB in order to avert some catastrophe, it might be permissible to kill the person but not the SEB. It has also been argued that the existence of SEBs might be bad for mere persons in ways that are unrelated to inviolability. For example, SEBs might have stronger claims to scarce resources, such as healthcare, than ordinary persons (Douglas 2011a).

Finally, a third response to concerns regarding the creation of SEBs has been to argue that, even if human neuroenhancements could create SEBs *and* this would be bad for ordinary persons, it may still be morally permissible, or even morally desirable, to create such beings. For example, Douglas (2011a, 2012), Persson (2012) and Wasserman (2012) argue that the costs to ordinary persons associated with the existence of SEBs might be compensated by other advantages or might be outweighed by goods enjoyed by the SEBs themselves. By contrast, Agar (2012a) has argued that the costs of MSNs for the unenhanced are unlikely to be compensated by other benefits.

Moral Desirability Neuroenhancements

Moral desirability enhancements (MDNs) aim to augment and (expectably) succeed in augmenting the moral desirability of a person's character traits, motives, or conduct (henceforth, collectively, "psychological features"). The moral desirability

[1]Each of these authors relies to some extent on an inductive inference, most fully spelled out by Agar (2012a): observed differences in mental capacity generate differences in moral status, and unobserved differences in mental capacity, between us and mentally enhanced beings, could be at least as great; thus we should expect that these too will give rise to differences in moral status. David Wasserman (2012) has recently offered an inductive argument militating in the opposite direction. He notes that the development of civilization, for example, from feudalism to contemporary social arrangements, has tended to elide distinctions in moral status so that we now accept fewer tiers of moral status than previously. If we assume that this is due to moral-epistemic progress, then we might infer that we are in the process of converging on a correct view of moral status that accepts very few tiers of moral status (perhaps simply a binary view that entities either have moral status or not). Such a view might render implausible the claim that mentally enhanced beings could have greater moral status than us.

of a psychological feature – often referred to simply as its moral goodness – corresponds to, or is at least correlated with, the degree to which there are typically agent-neutral moral reasons to promote that feature. It is to be distinguished from the moral worth or *moral praiseworthiness* of a psychological feature (also sometimes referred to using the term "moral goodness") which is instead standardly understood as the degree to which an agent merits moral praise for possessing that feature. Morally desirable traits are often also morally worthy, but they need not be.

Some psychological features may be *noninstrumentally* morally desirable. For example, Kant (1964) can be interpreted as arguing that a good will is noninstrumentally morally desirable, many have held that those character traits which qualify as virtues are noninstrumentally morally desirable, and many would also hold that respectful actions are, at least in one respect, noninstrumentally morally desirable. But character traits, motives, and actions can also be *instrumentally* morally desirable, because they produce outcomes that are themselves morally desirable. For example, breaking a promise, might be in one way instrumentally morally desirable when it results in lives being saved, pain being averted, or knowledge being gained. In what follows, moral desirability will be taken to include both instrumental and noninstrumental moral desirability. MDNs can thus be understood as interventions aim to increase and (expectably) succeed in increasing the total instrumental and noninstrumental value of a person's character traits, motives, or conduct.

Note that, on some views, the same interventions will qualify as both moral status neuroenhancements and moral desirability neuroenhancements. For example, on some Kantian views, possession of a good will is both a morally desirable character trait and what confers moral status on a being. Some such views might allow that improving a person's will via neuroenhancement – supposing that this were possible – could constitute both a MSN and a MDN.

Types of Moral Desirability Neuroenhancement

MDNs are unified by the kind of moral value that they augment: moral desirability. However, as already noted, they may differ in their target, with some targeting a person's conduct, others her motives, and others still her character traits. MDNs that target a person's character or character traits could be aptly described as enhancements of moral virtue since, on some accounts of virtue, the moral virtues are understood to be morally desirable character traits. However, moral virtue is also sometimes understood in other ways – for example, as referring to morally *praiseworthy* character traits.

MDNs can also be distinguished by the means they employ. As instances of neuroenhancement, all MDNs operate by directly modulating the brain states of an agent. However, we can distinguish them according to the way in which brain modulation is supposed to enhance the moral desirability of the agent's character,

motives, or conduct. Some MDNs might operate by directly altering an agent's mental *capacities or abilities* but without directly influencing any mental state. For example, Fröding (2011) suggests that certain cognitive neuroenhancements, such as those which mitigate cognitive biases, would in some cases facilitate the development of moral virtues. Similarly, one can imagine that an intervention which enhanced an agent's ability to vividly imagine the consequences of her actions might contribute to morally more desirable conduct.

These interventions would alter mental states only indirectly, by influencing the agent's mental capacities. Other MDNs might operate by directly influencing an agent's mental states, for example, by directly altering the agent's beliefs, desires, intentions, or emotions. Perhaps an intervention could enhance the moral desirability of an agent's conduct by augmenting her feelings of sympathy, her desire to help others, or her belief in the importance of moral requirements. Within the category of MDNs that directly influence an agent's mental states, we can distinguish between those that directly influence cognitive states (such as beliefs), conative states (such as desires and intentions), and affective states (such as emotions). As discussed below, some concerns that have been raised about MDNs are specific to MDNs that operate by directly altering mental states of certain kinds.

Scientific Prospects for Moral Desirability Neuroenhancement

Spence (2008) argues that, within psychiatry, some medications, such as those used to treat antisocial personality disorder, substance abuse, and psychosis, are already being used as MDNs in some circumstances. It might also plausibly be argued that anti-testosterone agents and other medications used to prevent sexual re-offending in some European and North American jurisdictions are intended to improve the moral desirability of the sex offender's conduct.

Most discussions of MDNs have, however, treated such enhancements as a prospect for the future rather than a contemporary phenomenon. To establish the plausibility of the claim that MDNs might become feasible or commonplace in the future, some have pointed to recent findings from moral psychology and neuroscience which are beginning to uncover the neural bases of morally significant character traits, motives, and behavior and, in some cases, to suggest possible interventions capable of modifying these (e.g., Savulescu and Persson 2012). For example, the neurotransmitter oxytocin appears to have significant affects on trust and cooperativeness between people (e.g., De Dreu et al. 2010, 2011; Kosfeld et al. 2005; Zak et al. 2004), and a number of widely used drugs, including the beta-blocker propranolol and selective serotonin reuptake inhibitors, a class of antidepressants, appear to have significant effects on human moral psychology (e.g., Crockett et al. 2008, 2010; Terbeck et al. 2012). Shook (2012) and Levy et al. (forthcoming) review some recent empirical findings relevant to future prospects for developing new MDNs.

Motives for Discussing the Morality of Moral Desirability Neuroenhancements

There have been two main motives for assessing the morality of MDNs. Some have assessed hypothetical MDNs as a way to test arguments that have been made regarding the morality of neuroenhancement or bioenhancement more generally (section introduction in ► Chap. 75, "Neuroenhancement" and ► Chap. 80, "Reflections on Neuroenhancement," ► Chap. 69, "Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues" ► Chap. 76, "Ethics of Pharmacological Mood Enhancement"). For example, Douglas (2008) presents certain MDNs as counterexamples to the claim that it is always morally impermissible to biologically augment one's capacities from an already normal level.

Other authors have discussed MDNs for more practical reasons: because they believe that there is a realistic prospect of MDNs being attempted in the future and wish to encourage, forestall or morally appraise such attempts or the technological developments that would enable them. For example, Ingmar Persson and Julian Savulescu have argued that MDNs might help humanity to escape its own destruction and should therefore be pursued, while John Harris, Robert Sparrow, Nick Agar, and others have raised serious moral concerns regarding MDNs.

The following two sections outline some of the main moral positions that have been taken on MDNs and the dominant arguments that have been made for and against those positions. As in the discussion of MSNs, arguments concerning the *possibility* of MDNs are discussed only insofar as they have figured in arguments concerning their morality.

Persson and Savulescu's Defense of an Imperative to Pursue the Development of MDNs

The most forthright and provocative defense of MDNs has been offered by Ingmar Persson and Julian Savulescu, who argue that there is "an urgent imperative to enhance the moral character of humanity" and to pursue research into moral neuroenhancements as a possible means to this end (2008, p. 162; see also 2010, 2011a, b, 2012). Persson and Savulescu argue that scientific progress, aided by cognitive neuroenhancement, is likely to bring it about that even small numbers of malevolent agents could cause great harm. For example, they suggest that scientific advances may allow a single scientifically trained individual to produce a designer pathogen capable of wiping out humanity, making it almost certain that humanity will indeed be destroyed. On the other hand, they argue, scientific progress will not have any benefits sufficiently great to make it rational to accept this risk of great harm, in part because it is generally easier to cause a harm than to cause a comparably large benefit and in part because it is rational to be somewhat more averse to harms than attracted to benefits. They maintain that:

> It will be bad for us that scientific knowledge continues to grow by traditional means, and even worse if this growth is further accelerated by biomedical or genetic enhancement of our cognitive capacities. For if an increasing percentage of us acquires the power to destroy a large number of us, it is enough if very few of us are malevolent or vicious enough to use this power for all of us to run an unacceptable increase of the risk of death and disaster. To eliminate this risk, cognitive enhancement would have to be accompanied by a moral enhancement which extends to all of us, since such moral enhancement could reduce malevolence. (2008, p. 166)

This moral enhancement, as they understand it, will consist in the acquisition of morally better (in the sense of more morally desirable) motivational dispositions. Persson and Savulescu argue that neither traditional forms of moral enhancement, such as moral education, nor cognitive enhancements (traditional or otherwise) are likely to produce the necessary moral enhancement. They also suggest that it is unlikely that the risks of great harm can be mitigated by holding back cognitive enhancement and scientific progress (2008, p. 173). They argue that we should thus pursue the possibility of moral neuroenhancements by, among other things, conducting scientific research that might facilitate their development (on this claim, see also Hughes 2006, 2011). This argument has attracted four main types of criticism.

Criticism 1: Misconstrual of the Risks and Benefits of Scientific Progress

One critical response has been to argue that Persson and Savulescu have overstated the risks or understated the benefits associated with further scientific progress. For example, Elizabeth Fenton maintains that, in characterizing the likely benefits of scientific progress as consisting in "only a *small* increase in an *already high* quality of life," Persson and Savulescu both underestimate the size of the likely benefits of scientific progress and misconstrue the likely starting point (2010, p. 149, her italics). She argues that the benefits of scientific progress would in many cases be enjoyed by individuals whose quality of life is poor, for example, due to disease, famine, or pollution, and for whom the benefits of such progress would be substantial. She also argues, appealing to Buchanan (2008), that further scientific progress might in fact be necessary to prevent significant decreases in quality of life (p. 150).

John Harris (2011) also offers a response of this second variety. He disputes Persson and Savulescu's claim that it is generally easier to harm than to benefit others by appealing to cases in which it seems easy to cause large benefits by making donations, initiating public health programs, providing vaccines, or preventing others from inflicting great harm (pp. 106–107). Suppose you know that someone is planning a mass shooting and have access to the gun they will use; you might then be able to save many lives by removing the bullets from this gun. Harris suggests that in this sort of case it is at least as easy for you to save the lives at stake as it is for the prospective shooter to end them.

Harris has also argued that cognitive neuroenhancement is itself a variety of moral neuroenhancement (2011, e.g., pp. 106, 110). The implication is that if future

technological progress is driven in part by cognitive neuroenhancement, we can expect that future people will also have more morally desirable character traits, motives, and conduct than us. This might somewhat mitigate the risk that new technologies will be used to harmful effect.

Persson and Savulescu offer a number of replies to claims that they have misconstrued the risk-benefit balance associated with further technological development, but perhaps the most important of these involves an appeal to the concept of ultimate harm: this is the harm of bringing it about that worthwhile life will never again exist. Persson and Savulescu argue that scientific progress will increase the risk of ultimate harm, and they take this to be an indefinitely bad outcome since, we have no way of knowing how much net value, in the form of continued worthwhile life, the ultimate harm prevents (e.g., 2011a, p. 442). On the other hand, they doubt that technological progress will similarly increase the likelihood of indefinitely good benefits.[2] Thus, they claim, we are in the position of weighing indefinitely bad harms against substantial but definite benefits, and faced with this choice it is rational to prefer that there is no further technological progress.

Criticism 2: Misuse of Moral Enhancement Technologies

Persson and Savulescu are concerned that the products of technological development might be misused to devastating effect. Some authors have responded by noting that the moral enhancement technologies whose development they call for, or the scientific information that would need to be acquired in order to develop them, could themselves be misused by malevolent agents (Ehni and Aurenque p. 231–2; Fenton 2010; Harris 2011, p. 108; Sparrow forthcoming a, b). Presumably, if we were technologically capable of bringing it about that people have *more* morally desirable character traits, motives, or conduct, we could also bring it about that people have *less* desirable traits, motives, or conduct, and some agents might well be more interested in decreasing moral desirability than increasing it. For example, we can imagine that a ruthless businessman or soldier might want to suppress pangs of conscience and that this might bring about morally undesirable motives and conduct.

Persson and Savulescu's response to these concerns has been to concede that technologies which enable MDNs could be misused but maintain that pursuit of MDNs is nevertheless our only hope of avoiding ultimate harm (Persson and Savulescu 2011a, p. 443).

[2]Persson and Savulescu (2011b, p. 4) do allow that one might contribute to an indefinitely large benefit by preventing someone else from causing ultimate harm, something that technological progress might allow. However, they argue that, in that sort of case, one would not be *guaranteeing* an indefinitely large benefit, because someone else (or a natural disaster) might cause ultimate harm. By contrast, when one causes ultimate harm, one does guarantee an indefinitely bad harm.

Criticism 3: Implausible Implications

Persson and Savulescu have also been criticized for being committed to intuitively implausible views. For example, Fenton (2010) argues that they are committed to the view that "*all* forms of scientific progress are instrumentally bad for humans overall" (p. 149, her italics). Their response is to argue that there has been a turning point. Before that point, scientific progress could be expected to produce net benefits; afterwards, it could be expected to produce net harms. They suggest that this turning point occurred when "science and technology put in the hands of humans the means of destroying or seriously damaging forever the conditions of sentient life on this planet" and speculate that this point may have come in the mid-twentieth century (Persson and Savulescu 2011a, p. 441). Meanwhile, Harris (2011) argues that Persson and Savulescu are committed to the view that it would be in one way desirable to retard the cognitive powers of current and future humans, thus slowing scientific progress.

Criticism 4: Moral Neuroenhancement Won't Help

A fourth criticism of Persson and Savulescu's argument maintains that pursuit of moral enhancement would do little to diminish the risk of ultimate harm (Harris 2011, pp. 109–10). Since the risks that worry Persson and Savulescu are risks that could be created by lone malevolent agents, and since those agents are also likely to be able to avoid undergoing moral enhancement, it may seem doubtful whether pursuit of moral enhancement would do much to reduce these risks.

In response, Persson and Savulescu (2012) argue that technological progress can exacerbate risks of ultimate harm not only by allowing lone agents to wreak great havoc but also by exacerbating the negative effects of widespread but mundane moral failings. For example, they argue that climate change may cause ultimate harm and that it can be attributed to a combination of technological development and human moral failings. Moral neuroenhancement might be capable of substantially reducing climate change even if not universally undergone. Thus, they suggest, MDNs might reduce some risks of ultimate harm even if they would fail to reduce others.

Other Defenses of MDNs

Persson and Savulescu are interested in whether humanity falls under an imperative to pursue or promote the development of technologies that would enable MDNs. However, there are many other moral questions that we might ask regarding the morality of MDNs. For example, we might ask whether there is some weaker moral conclusion that could be made in favor of pursuing the development of MDNs: even if humanity falls under no imperative to promote MDNs, it might nevertheless be morally permissible, or even morally desirable, for it to do so. We could also

move away from questions about what humanity as a whole ought to do to questions about what individual people, or groups of people, ought to do. Again, even if humanity falls under no imperative to promote MDNs, it might nevertheless be permissible, desirable, or obligatory for individual people or groups of people to promote MDNs, to undergo MDNs themselves, or to encourage or require others to do so.

There is thus scope for weaker defenses of MDN, and a number of such defenses have been offered. For example, Thomas Douglas (2008) has argued that it would often be morally permissible for individuals to voluntarily engage in MDNs that expectably bring it about that they have more morally desirable motives than they would otherwise have had. More recently, David DeGrazia (2012b) has argued that moral desirability enhancements of motives or behavior would themselves be morally desirable, under certain idealizing assumptions. There have also been weaker defenses of specific means to MDN. For example, Douglas (2011b) argues that it could be morally permissible to enhance the moral desirability of one's motives or conduct via *directly modulating one's emotions* – that is, via means which, once set in train, modulate one's emotions without requiring the engagement of one's deliberative capacities.

It would be relatively uncontroversial that individuals have moral reasons to enhance the moral desirability of their character, motives, and conduct, and that doing so is in one way morally desirable. These may even be analytic truths. Moreover, MDNs appear to be immune to many of the general concerns that have been raised about neuroenhancements (or bioenhancements more generally). These concerns have often focused on ways in which neuroenhancements undergone by some individuals might harm or wrong others, for example, by placing them at an unfair competitive disadvantage or undermining commitments to solidarity or equality. MDNs are unusual among the main types of neuroenhancements that have been discussed heavily in recent literature in that they might plausibly be expected to advantage, rather than disadvantage, others, on balance.

Nevertheless, some significant concerns have been raised regarding the permissibility and desirability of undergoing MDNs or certain kinds of MDNs. Some of these are general concerns about enhancing the moral desirability of our characters, motives, and conduct, whether through *neuro*enhancement or more traditional means. In this category are general skepticism about the possibility of moral improvement and concerns about whether we have adequate means of resolving disagreement and uncertainty about what character traits, motives, and conduct are morally desirable and why. Debate regarding MDNs has elicited significant moral discussion on these points (Jotterand 2011; Schaefer 2011; Shook 2012; Wasserman 2011), but these concerns apply to traditional means of augmenting moral desirability as well as MDNs. Other concerns are general concerns about neuroenhancement that would apply to nonmoral neuroenhancements as much as to MDNs. In this category are concerns regarding the unnatural means or hubristic motives that neuroenhancement is said to involve (Kass 2003; Sandel 2007). Other concerns are, however, specific to MDNs – they would not apply equally to traditional means of enhancing moral desirability or to other kinds of

neuroenhancement. The remainder of this section outlines three dominant concerns in this category as well as some responses that have been offered to them by those who wish to defend at least some MDNs.

Concern 1: Restriction of Freedom

One concern that has been raised regarding MDNs, or certain MDNs, is that such enhancements might restrict freedom or autonomy. Thus, for example, John Harris (2011) argues that we might have reason to abstain from MDNs because they would restrict our freedom to perform morally undesirable actions or to have morally undesirable motives (see also Ehni and Aurenque 2012, p. 233 and, for a more general discussion of the effects of neuroenhancement on autonomy, Bublitz and Merkel 2009).

A similar thought underpins the well-known free will defense of theism. The free will defense maintains that the existence of an omnipotent, omniscient, and benevolent God is consistent with the presence of evil because evil is a consequence of our possessing the freedom to do evil, which is, all things considered, good. Though the freedom to do evil possesses the instrumental disvalue of allowing evildoing, it also possesses some other, greater value. It is no surprise, then, that an omnipotent, omniscient, and benevolent God would have allowed evil to exist.

If the freedom to do evil is all-things-considered valuable despite its very great instrumental disvalue, one might expect that the more general freedom to have morally undesirable character traits or motives or to act in morally undesirable ways (henceforth, the freedom to be immoral) is also good despite its instrumental disvalue. Perhaps, then, MDNs will be all-things-considered disvaluable whenever they restrict this freedom.

Two main types of response have been made to this suggestion. The first has been to argue that, even where MDNs do restrict freedom, it might nevertheless be morally permissible or all-things-considered morally desirable to undergo such MDNs. The second has been to deny that all MDNs would restrict freedom, thus limiting the concern about freedom to a subset of MDNs.

Responses of the first type parallel a standard response to the free will defense of theism. That response holds that, in many cases, it seems preferable to sacrifice some freedom to do evil in order to prevent evil than to tolerate both the freedom to do evil and the associated evil. If I witness one person about to murder another, it seems that I should intervene to prevent the murder even though this involves restricting the prospective murder's freedom to do evil. Similarly, it seems that a would-be murderer should restrict his own freedom to do evil, thus preventing murder, if he is in a position to do so. The obvious way of explaining this conclusion is to suppose that, in at least some cases, any disvalue associated with restricting one's freedom to do evil is outweighed by the value of doing so. Similarly, it has been argued that there will be MDNs where the disvalue of any loss in freedom to be immoral is outweighed by the

value of increasing the moral desirability of one's motives or conduct (DeGrazia 2012b; Douglas 2008, 2011b; Savulescu et al. forthcoming; Savulescu and Persson 2012).

Responses in the second category sometimes begin by noting that concerns about the freedom-reducing effect of MDNs presuppose compatibilism – namely, the view that freedom is consistent with one's motives and conduct being causally determined. For if compatibilism were false, we could be free only if we were causally undetermined. In that case, we would either already be completely unfree, because we are causally determined, in which case MDNs could not *reduce* our freedom, or we are free only because we are, at least some of the time, immune from causal forces, in which case MDNs, which operate through causal channels, could not affect us in circumstances where we are currently free (Blackford 2010; DeGrazia 2012b; Savulescu et al. forthcoming; Savulescu and Persson 2012).

Thus, it is argued, concerns about the freedom-reducing effect of MDNs will have to be grounded on compatibilism. But on compatibilist views of freedom, whether MDNs diminish freedom is likely to depend on precisely how they operate.

We can think of compatibilists as distinguishing between two aspects of our psychological life: a true or authentic self (which may be identified with higher order desires, or higher order desires that are not the result of inauthentic influences) and a brute self. Motives or actions are then thought of as free if and only if they are part of or flow from the true self.

On this picture, MDNs could reduce our freedom to be immoral if they operate by strengthening the influence of the brute self vis-à-vis the true self. Consider an individual who, whenever approached by poverty-relief charities, feels a conflict between, on the one hand, a firm commitment, formed through introspective reflection, to the view that she owes nothing to the poor, and on the other hand, feelings of sympathy towards the poor which are the result of extraneous social pressures and which she experiences as unwanted. Suppose further that this individual, under social pressure, undergoes an intervention that directly augments her feelings of sympathy. Such an intervention might perhaps qualify as a MDN, and it will also, on many compatibilist views, restrict her freedom. A proponent of these views would deem this agent's feelings of sympathy, both before and after the intervention, to be aspects of her brute self, while her moral commitments regarding charity are an aspect of her true self. The intervention might thus plausibly be construed as one in which the influence of the agent's brute self is expanded at the expense of the agent's true self. One might expect that, following the intervention, a greater proportion of this agent's motives and conduct would be unfree, where that implies that they are part of or flow from the brute, rather than the true, self.

However, defenders of MDNs have argued that, just as we can paint a compatibilist picture of a MDN that restricts the freedom to be immoral, so too we can paint a picture of one that operates by increasing the freedom to be moral. They argue that, on any plausible compatibilist account of freedom, there will be some ways in which the conduct of some people is already unfree – because it flows from aspects of our brute selves. MDNs could operate precisely by attenuating the

influence of these aspects of our brute selves (Douglas 2008, 2011b; Savulescu et al. forthcoming). Thus, suppose that, when an agent is approached by worthy poverty-relief charities, she feels a pull between a reflectively formed commitment to the view that poverty-relief is morally required and a hoarding instinct not to part with any of one's property that is unwanted and the result of external factors. In this case, many compatibilist views will deem the moral commitment to be part of the true self and the hoarding instinct to be part of the brute self. Thus, it will be plausible that an MDN that attenuated the hoarding instinct would increase the agent's freedom.

Concern 2: Superficiality

A second concern that has been raised regarding MDNs is targeted specifically at Douglas' defense of MDNs that increase the moral desirability of one's motives or conduct by directly modulating emotions (call such interventions "Emotional MDNs"). John Harris worries that an intervention which operates in this way "cannot be moral enhancement properly so called at all". Indeed, he maintains that "the notion of moral behaviour has been attenuated to a vanishing point" once one claims that such behavior could be produced by directly altering emotions (2012a, p. 6); "tinkering with the emotions is not a form of moral enhancement at all. It is more like the threat of punishment: it may make immoral behaviour less likely, but it does not enhance morality" (2012b, pp. 3–4; see also Harris and Chan 2010; Sparrow forthcoming a, b).

One way of reading these passages would see them as asserting that Emotional MDNs are metaphysically, or at least physically, impossible. They violate some metaphysical or physical law. (Doubts regarding the possibility of MDNs more generally have also been raised by Ehni and Aurenque 2012, p. 232; Jotterand 2011; and Shook 2012.) However, it is difficult to see how this charge could be sustained. Emotions are mental states; mental states are normally taken to be either constitutively or causally dependent on brain states, and brain states are susceptible to direct modulation. It seems both metaphysically and physically possible to alter the emotions by directly influencing brain states. It also seems metaphysically and physically possible for alterations in one's emotions to alter the moral desirability of one's motives or conduct. For example, gratuitous killing of innocents in war is morally undesirable, and it is plausible that differences in the extent to which a solider experiences emotions such as an uncontrollable rage might influence whether or not that soldier engages in such morally undesirable conduct.

Perhaps a more plausible way of interpreting Harris' passages would see them as conceding that direct modulation of emotions could increase the moral desirability of one's motives or conduct but maintaining that it could not produce some deeper kind of moral improvement for which the term "moral enhancement" should be reserved. Fabrice Jotterand (2011) makes a similar point, arguing that "moral neuroenhancement is unlikely to morally enhance people in the true meaning of the word" (p. 8), as does Robert Sparrow (forthcoming b), who suggests that

"while there is indeed evidence that certain pharmaceutical and neuro-scientific interventions can alter dispositions and behaviour in ways that we may be inclined to morally evaluate positively, this falls well short of constituting 'moral bioenhancement' in any interesting sense." These authors might be read as suggesting that the direct modulation of emotions could not produce *deep* moral improvement, where that might consist, for example, in an increase in the moral *worth* or *virtue* of one's motives or conduct.

It would not, of course, follow from this that Emotional MDNs are absolutely morally undesirable in any way. To claim that Emotional MDNs fail to produce deep moral improvement is to claim that they lack some desirable feature, not that they possess some absolutely undesirable one. However, if Emotional MDNs fail to produce deep moral improvement, it might follow that they are *less* morally desirable than certain other ways of enhancing the moral desirability of our conduct that do produce deep moral improvement. Harris (2011, 2012a, b) can be read as maintaining that, while Emotional MDNs fail to produce deep moral improvement, more traditional, deliberative means of enhancing moral desirability, such as introspective reflection and engagement with literature, do produce such improvement and should thus generally be adopted in preference to MDNs.

Harris buttresses his claim that Emotional MDNs could not produce deep moral improvement by drawing a parallel between "being moral," in the deeper sense in which he is interested, and "being scientific":

> [o]ne can accidentally discover something of scientific importance, but one cannot be scientific, one cannot do science, accidentally. Doing science is a deliberative and disciplined process. It involves, for example, doing things like formulating and testing a hypothesis and looking for disconfirmatory evidence as well as for confirmatory evidence. . .Being moral is like being scientific. (Harris 2012a, p. 6)

He suggests that, unlike traditional means of enhancing moral desirability, Emotional MDNs could bring about morally desirable motives and conduct only by accident. One might ground an argument for this view on the observation that traditional means of enhancing moral desirability frequently operate by enhancing an agent's moral understanding: her understanding of what morality requires and why, both in general and on particular occasions on which one must make a morally significant decision. This moral understanding is an all-purpose tool that helps the agent to be motivated and act in morally desirable ways in many circumstances. If an agent understands what morality requires and why, she will be well placed to be motivated and act in morally desirable ways in almost any circumstance.

On the other hand, Harris (2012a, b) suggests, Emotional MDNs would not operate by enhancing the agent's moral understanding. Others have made similar claims. For example, Fabrice Jotterand (2011) argues that "[w]hile the manipulation of moral emotions might change the behavior of an individual, it does not provide any content, for example, norms or values to guide one's behavioral response" (p. 6, see also p. 8). Similarly, Robert Sparrow (forthcoming a) suggests that "[i]t is hard to see how any drug could alter our beliefs in such a way as to track the reasons we have to act morally" and that "[s]omeone who reads Tolstoy arguably learns to be less

judgemental and in doing so develops greater understanding: someone who takes a pill has merely caused their sentiments to alter." If these authors are correct, Emotional MDNs (and perhaps MDNs more generally) do not augment moral understanding. Instead, it might be argued, Emotional MDNs would typically work by removing some relatively straightforward emotional barrier to morally desirable motivation or conduct. The most obvious examples of such obstacles might include aggressive impulses, strongly xenophobic sentiments, or a callousness or ruthlessness towards the suffering of others. But note that these are not *universal* barriers to moral desirable motivation and conduct. For example, one can imagine circumstances in which aggressive impulses or ruthlessness would produce morally desirable motives and conduct; aggressiveness might do so when one is fighting a just war, or perhaps when one is confronted with one person assaulting another on the street, and ruthlessness might do so when one is a politician engaged in diplomatic negotiations (Douglas 2008; Savulescu et al. forthcoming; Wasserman 2011, forthcoming). Thus, it might be thought that Emotional MDNs would produce only accidental or unreliable morally desirable motives and conduct.

There is little literature responding to Harris' concern about the superficiality of Emotional MDNs nor to the suggestion that Emotional MDNs could produce no more than accidental morally desirable motives and conduct. However, some defenders of MDN have suggested that such neuroenhancements could operate by attenuating emotional barriers to sound moral deliberation (Douglas 2008, forthcoming) or by bringing it about that we are more like those existing people who we regard as especially moral (Savulescu and Persson 2012; Savulescu et al. forthcoming), and one might expect that where they operate in this way they will, like more traditional forms of moral enhancement, increase moral understanding. In addition, Wasserman (2011) has argued that, even if an Emotional MDN initially had no positive effect on moral understanding, we might expect the frequent presence of pro-moral emotions and the morally desirable motives or conduct to lead to a development in moral understanding over time. This parallels the Aristotelian point that one comes to know the good by being good (Burnyeat 1980).

Concern 3: Misfiring

A third concern that can be raised regarding MDNs maintains that attempts at MDNs are likely to misfire, bringing about moral deteriorations rather than improvements. This is not a concern about successful MDNs but is rather the concern that actual attempts at MDN are likely to be unsuccessful.

Harris (2011) advances this concern by noting that "the sorts of traits or dispositions that seem to lead to wickedness or immorality are also the very same ones required not only for virtue but for any sort of moral life at all" (p. 104). He infers from this that the sorts of psychological alterations required for MDN would involve not the wholesale elimination or dramatic amplification of particular dispositions, but rather a kind of fine-tuning of our dispositions (see also Jotterand 2011, p. 7; Wasserman 2011). However, he argues that the disposition-modifying neurotechnologies that we are

actually likely to have available to us will be rather blunt: it will be difficult to determine or even predict their effects precisely or to target the specific dispositions that we wish to shape. Thus, he suggests, it is unlikely that any attempt to use these technologies will succeed in bringing about an improvement in the moral desirability of our motives or conduct.

Nicholas Agar (2010, 2012b, under review) puts forward a more limited variant of this concern. He argues that attempted MDNs may have good chances of success when they aim only to correct subnormal levels of moral desirability, bringing an individual within the normal range, but that they are likely to misfire when they aim to produce levels of moral desirability above the normal range. Subnormal moral desirability is often the result of relatively isolated and easily identified defects such as, for example, the deficient empathy that characterizes psychopathy. Agar speculates that these defects could relatively safely be corrected. However, he argues that, to attain supranormal levels of moral desirability, we would need to simultaneously augment or attenuate several different dispositions in a balanced way. This, he claims, will be very difficult. One might suppose that this difficulty derives from three sources. First, there may be moral uncertainty – uncertainty regarding what character traits, motives, or conduct qualify as supranormally morally desirable. There would be significant disagreement on this matter and that this disagreement may be both evidence for and a source of uncertainty (see, e.g., Schaefer 2011). Second, there may be empirical uncertainty about what changes would need to be wrought in a given individual to realize the putatively desirable traits. Third, there might, for reasons suggested by Harris (2011), be practical difficulties in realizing these transformations using technologies that we have available to us. Taken together, these difficulties might create a serious risk that attempts to bring about supranormal MDNs will fail.

The main response made to these concerns by defenders of MDN has been concessive. Defenders of MDNs have acknowledged both that (1) in many cases, complex and subtle interventions would be needed in order to enhance moral desirability and that (2) this creates a risk that attempted MDNs will fail, perhaps instead resulting in moral deterioration (Douglas 2011b; Savulescu et al. forthcoming). However, some doubt has been cast on the thought that this concern would always count decisively against attempting MDNs. For example, Douglas (2011b) notes that there are other areas – such as clinical psychiatry – where we often also use rather blunt biological interventions as part of efforts to achieve subtle and multidimensional psychological changes. Yet in that area we normally think that attempting some interventions can be permissible and desirable if undergone cautiously, keeping open the option of reversing the intervention if it misfires. He suggests a similar approach might be justified in relation to MDNs.

Further Questions

Moral debate on MDNs has focused primarily on (1) the morality of promoting the development of technologies that would enable MDNs, as in the case of discussion

surrounding Persson and Savulescu's proposals, and on (2) the morality of voluntarily undergoing MDNs oneself, as in the case of discussion surrounding Douglas' and DeGrazia's proposals. Further questions might be asked regarding the morality of requiring or encouraging others to undergo such neuroenhancements. For example, one might wonder whether it would ever be permissible for a state to impose MDNs on its citizens, for courts to impose MDNs on convicted offenders, for parents to impose MDNs on their children, or for employers to impose MDNs on their employees. Such uses of MDNs would raise, in addition to the concerns discussed above, concerns regarding, inter alia, coercion, manipulation, and domination. For example, Sparrow suggests that imposing moral enhancements on others might, even if benevolently motivated, constitute an objectionable form of domination (forthcoming; under review; see also Bublitz and Merkel 2009). Walker (2009) and Blackford (2010) have responded to similar concerns as they might be raised in relation to the imposition of MDNs on one's children or children-to-be.

Perhaps the context in which coercive use of MDNs is most likely to gain acceptance is that of criminal justice. Arguably, institutions of criminal justice are, at least insofar as they offer criminal rehabilitation programs, already engaged in a kind of moral enhancement. In addition, concerns about moral disagreement might be regarded as least serious in this context; criminal conduct is conduct that, if not universally taken to be morally undesirable, is at least widely accepted as conduct which the state and citizenry may legitimately seek to prevent. Concerns regarding coercion might also be regarded as weaker in the context of criminal justice, since a degree of coercion is also already accepted in that area: the paradigmatic criminal sanction, incarceration, is itself highly coercive. Nevertheless, it might be argued that forced MDNs in criminal justice would be more problematic than either incarceration or psychological rehabilitation programs, perhaps because more intrusive or manipulative. There is a burgeoning literature on this topic (Bomann-Larsen 2011; Bublitz and Merkel 2012; Rosati 1994; Ryberg and Petersen 2011; Shaw 2012; Vincent 2012).

Questions might also be asked regarding the morality of selecting for morally desirable character traits in one's children, for example, on the basis of genetic information acquired through preimplantation genetic diagnosis. Moral enhancement and moral neuroenhancement have been understood in this chapter to involve augmenting the morality of some particular individual or her traits. On this understanding, selecting for moral traits in one's children would not qualify as moral neuroenhancement, or indeed moral enhancement, since it instead involves bringing one (expectably) more moral individual into existence in preference to another (expectably) less moral individual. However, some of the same issues arise in cases of selection as in cases of neuroenhancement discussed here, and there is some literature exploring these. For example, Faust (2008) argues that if parents were able to select for morally desirable traits in their future children, it would be desirable, and probably obligatory, for them to do so. Walker (2009) argues that there is no principled reason not to select for moral virtues in one's children, unless there is a principled reason not to engage in

selection at all.[3] For criticism of these claims, see Agar (2010), Andreadis (2010), Arnhart (2010), and Bronstein (2010).

Cross-References

- ▶ Ethics of Pharmacological Mood Enhancement
- ▶ Neuroenhancement
- ▶ Reflections on Neuroenhancement
- ▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Agar, N. (2010). Enhancing genetic virtue? *Politics and the Life Sciences, 29*(1), 73–75. doi:10.2990/29_1_73.

Agar, N. (2012a). Why is it possible to enhance moral status and why doing so is wrong? *Journal of Medical Ethics*. doi:10.1136/medethics-2012-100597. http://jme.bmj.com/content/early/2012/08/23/medethics-2012-100597.short

Agar, N. (2012b). A question about defining moral bioenhancement. *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101153.

Agar, N. (Under review). Against moral bioenhancement.

Andreadis, A. (2010). The tempting illusion of genetic virtue. *Politics and the Life Sciences, 29*(1), 76–78. doi:10.2990/29_1_76.

Arnhart, L. (2010). Can virtue be genetically engineered? *Politics and the Life Sciences, 29*(1), 79–81. doi:10.2990/29_1_79.

Blackford, R. (2010). Genetically engineered people. *Politics and the Life Sciences, 29*(1), 82–84. doi:10.2990/29_1_82.

Bomann-Larsen, L. (2011). Voluntary rehabilitation? On neurotechnological behavioural treatment, valid consent and (in)appropriate offers. *Neuroethics*. doi:10.1007/s12152-011-9105-9. http://www.springerlink.com/content/314k2722666347j5/

Bronstein, J. (2010). Objecting to the genetic virtue program. *Politics and the Life Sciences, 29*(1), 85–87. doi:10.2990/29_1_85.

Bublitz, J. C., & Merkel, R. (2009). Autonomy and authenticity of enhanced personality traits. *Bioethics, 23*(6), 360–374. doi:10.1111/j.1467-8519.2009.01725.x.

Bublitz, J. C., & Merkel, R. (2012). Crimes against minds: On mental manipulations, Harms and a human right to mental self-determination. *Criminal Law and Philosophy*. doi:10.1007/s11572-012-9172-y.

Buchanan, A. (2008). Enhancement and the ethics of development. *Kennedy Institute of Ethics Journal, 18*(1), 1–34.

Buchanan, A. (2009). Moral status and human enhancement. *Philosophy & Public Affairs, 37*(4), 346–381.

Buchanan, A. (2011). *Beyond humanity? The ethics of biomedical enhancement*. Oxford: Oxford University Press.

[3]Elster (2011) and Douglas and Devolder (forthcoming) have also defended ethical principles which imply that parents would typically have significant moral reasons to engage in such selection.

Buchanan, A. (2012). Still unconvinced, but still tentative: A reply to DeGrazia. *Journal of Medical Ethics, 38*(3), 140–141.

Burnyeat, M. F. (1980). Aristotle on learning to be good. In A. O. Rorty (Ed.), *Essays on Aristotle's ethics* (pp. 69–92). Berkeley: University of California Press.

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science, 320*, 1739.

Crockett, M. J., Clark, L., Hauser, M., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107*, 17433–17438.

Darwall, S. L. (2006). *The second-person standpoint: Morality, respect, and accountability*. Cambridge, MA: Harvard University Press.

De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Van Kleef, G. A., Baas, M., Ten Velden, F. S., Van Dijk, E., & Feith, S. W. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science, 328*(5984), 1408–1411. doi:10.1126/science.1189047.

De Dreu, C. K. W., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences, 108*(4), 1262–1266. doi:10.1073/pnas.1015316108.

DeGrazia, D. (2012a). Genetic enhancement, post-persons and moral status: A reply to Buchanan. *Journal of Medical Ethics, 38*(3), 135–139. doi:10.1136/medethics-2011-100126.

DeGrazia, D. (2012b). Moral enhancement, freedom, and what we (should) value in moral behavior. *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101157.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*(3), 228–245.

Douglas, T. (2011). Human enhancement and supra-personal moral status. *Philosophical Studies*, forthcoming in print. doi:10.1007/s11098-011-9778-2. http://www.springerlink.com/index/10.1007/s11098-011-9778-2

Douglas, T. (2011). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics*. doi:10.1111/j.1467-8519.2011.01919.x.

Douglas, T. (2012). The harms of status enhancement could be compensated or outweighed: A response to Agar. *Journal of Medical Ethics*. doi:10.1136/medethics-2012-100835. http://jme.bmj.com/content/early/2012/10/08/medethics-2012-100835

Douglas, T. (2013). Enhancing moral conformity and enhancing moral worth. Neuroethics. doi:10.1007/s12152-013-9183-y.

Douglas, T., & Devolder, K. (Forthcoming). Procreative altruism: Beyond individualism in reproductive selection. *Journal of Medicine and Philosophy*.

Ehni, H.-J., & Aurenque, D. (2012). On moral enhancement from a habermasian perspective. *Cambridge Quarterly of Healthcare Ethics, 21*(02), 223–234. doi:10.1017/S0963180111000727.

Elster, J. (2011). Procreative beneficence – Cui Bono? *Bioethics, 25*(9), 482–488. doi:10.1111/j.1467-8519.2009.01794.x.

Faust, H. S. (2008). Should we select for genetic moral enhancement? A thought experiment using the MoralKinder (MK+) haplotype. *Theoretical Medicine and Bioethics, 29*(6), 397–416.

Fenton, E. (2010). The perils of failing to enhance: A response to Persson and Savulescu. *Journal of Medical Ethics, 36*(3), 148–151. doi:10.1136/jme.2009.033597.

Fröding, B. E. E. (2011). Cognitive enhancement, virtue ethics and the good life. *Neuroethics, 4*(3), 223–234. doi:10.1007/s12152-010-9092-2.

Fukuyama, F. (2004). Transhumanism. *Foreign Policy, 144*(September), 42–43.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*(2), 102–111. doi:10.1111/j.1467-8519.2010.01854.x.

Harris, J. (2012). What it's like to be good. *Cambridge Quarterly of Healthcare Ethics*. doi:10.1017/S0963180111000867.

Harris, J. (2012). "Ethics is for bad guys!" putting the "moral" into moral enhancement. *Bioethics*. doi: 10.1111/j.1467-8519.2011.01946.x.

Harris, J., & Chan, S. (2010). Moral behavior is not what it seems. *Proceedings of the National Academy of Sciences, 107*(50), E183.
Hughes, J. (2006). Virtue engineering: Applications of neurotechnology to improve moral behavior, presented at *Transvision*, Helsinki. http://ieet.org/index.php/IEET/more/hughes20071120
Hughes, J. (2011). After happiness, cyborg virtue. *Free Inquiry, 32*(1).
Jotterand, F. (2011). "Virtue engineering" and moral agency: Will post-humans still need the virtues? *The American journal of bioethics: AJOB Neuroscience, 2*(4), 3–9. doi:10.1080/21507740.2011.611124.
Kant, I. (1964). *Groundwork of the metaphysic of morals*. 1st Harper torchbook edition (trans: Paton H.J.). New York: Harper & Row.
Kass, L. R. (2003). Ageless bodies, happy souls. *The New Atlantis, 1*, 9–28.
Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(2), 673–676.
Levy, N., Douglas, T., Kahane, G., Terbeck, S., Cowen, P. J., Hewstone, M., & Savulescu, J. (Forthcoming). Are you morally modified? The moral effects of widely used pharmaceuticals. *Philosophy*, *Psychiatry*, & *Psychology*.
McMahan, J. (2009). Cognitive disability and cognitive enhancement. *Metaphilosophy, 40*(3–4), 582–605. doi:10.1111/j.1467-9973.2009.01612.x.
Persson, I. (2012). Is Agar biased against "post-persons"? *Journal of Medical Ethics*. doi:10.1136/medethics-2012-100836. http://jme.bmj.com/content/early/2012/10/08/medethics-2012-100836.
Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy, 25*(3), 162–177.
Persson, I., & Savulescu, J. (2010). Moral transhumanism. *The Journal of Medicine and Philosophy, 35*(6), 656–669. doi:10.1093/jmp/jhq052.
Persson, I., & Savulescu, J. (2011a). The turn for ultimate harm: A reply to Fenton. *Journal of Medical Ethics, 37*(7), 441–444. doi:10.1136/jme.2010.036962.
Persson, I., & Savulescu, J. (2011b). Getting moral enhancement right: The desirability of moral bioenhancement. *Bioethics*. doi:10.1111/j.1467-8519.2011.01907.x. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8519.2011.01907.x/abstract
Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.
Rawls, J. (2001). *Justice as fairness: A restatement*. Cambridge, MA: Harvard University Press.
Rosati, C. S. (1994). A study of internal punishment. *Wisconsin Law Review, 1994*, 123–1579.
Ryberg, J., & Petersen, T. S. (2011). Neurotechnological behavioural treatment of criminal offenders – A comment on Bomann-Larsen. *Neuroethics*, 1–5. doi:10.1007/s12152-011-9146-0.
Sandel, M. (2007). *The case against perfection: Ethics in the age of genetic engineering*. Cambridge, MA: Harvard University Press.
Savulescu, J. (2009). The human prejudice and the moral status of enhanced beings: What do we owe the Gods. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement*. Oxford: Oxford University Press.
Savulescu, J., & Persson, I. (2012). Moral enhancement, freedom, and the God machine. *The Monist, 95*(3), 399–421.
Savulescu, J., Douglas, T., & Persson, I. (Forthcoming). Autonomy and the ethics of biological behaviour modification. In *Towards Bioethics in 2050*: *International Dialogues*.
Schaefer, G. O. (2011). What is the goal of moral engineering? *The American journal of bioethics: AJOB Neuroscience, 2*(4), 10–11. doi:10.1080/21507740.2011.620593.
Shaw, E. (2012). Direct brain interventions and responsibility enhancement. *Criminal Law and Philosophy*, 1–20. doi:10.1007/s11572-012-9152-2.
Shook, J. R. (2012). Neuroethics and the possible types of moral enhancement. *The American journal of bioethics: AJOB Neuroscience, 3*(4), 3–14. doi:10.1080/21507740.2012.712602.

Spence, S. A. (2008). Can pharmacology help enhance human morality? *The British Journal of Psychiatry: The Journal of Mental Science, 193*, 179–180. doi:193/3/179.

Sparrow, R. (Forthcoming a). (Im)moral technology? Thought experiments and the future of "mind control". In A Akabayashi (Ed.), *Towards bioethics in 2050*: *International dialogues*.

Sparrow, R. (Forthcoming b). Better living through chemistry? A reply to Savulescu and Persson on "moral enhancement". *Journal of Applied Philosophy*

Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Cowen, P., & Hewstone, M. (2012). Beta-adrenergic blockade reduces implicit negative racial bias. *Psychopharmacology, 222*(3), 419–424.

Vincent, N. (2012). Restoring responsibility: Promoting justice, therapy and reform through direct brain interventions. *Criminal Law and Philosophy* 1–22. doi:10.1007/s11572-012-9156-y.

Walker, M. (2009). Enhancing genetic virtue. *Politics and the Life Sciences, 28*(2), 27–47. doi:10.2990/28_2_27.

Wasserman, D. (2011). Moral betterness and moral enhancement Presented at the Uehiro-Carnegie Conference 2011, New York.

Wasserman, D. (2012). Devoured by our own children: The possibility and peril of moral status enhancement. *Journal of Medical Ethics.* doi:10.1136/medethics-2012-100843. http://jme.bmj.com/content/early/2012/09/27/medethics-2012-100843.

Wasserman, D. (Forthcoming). When bad people do good things: Will moral enhancement make the world a better place? *Journal of Medical Ethics*.

Wikler, D. I. (2009). Paternalism in the age of cognitive enhancement: Do civil liberties presuppose roughly equal mental ability. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (pp. 341–355). Oxford: Oxford University Press.

Wilson, J. (2007). Transhumanism and moral equality. *Bioethics, 21*(8), 419–425.

Wilson, J. (2012). Persons, post-persons and thresholds. *Journal of Medical Ethics, 38*(3), 143–144. doi:10.1136/medethics-2011-100243.

Zak, P. J., Kurzban, R., & Matzner, W. T. (2004). The neurobiology of trust. *Annals of the New York Academy of Sciences, 1032*, 224–227.

Reflections on Neuroenhancement

80

Walter Glannon

Contents

Abstract

Neuroenhancement through psychopharmacology can improve a range of cognitive and affective functions involved in practical and moral reasoning and decision-making. In this chapter, the claim that this type of enhancement would adversely alter human nature will be examined, and it will be argued that the claim is unfounded because it is based on questionable assumptions. In addition, some of the neurobiological risks associated with cognition-enhancing drugs will be considered. It will be pointed out that answers to questions about risk can only be provided after a sufficient number of prospective controlled studies of the drugs' effects have been conducted. On the basis of studies conducted thus far, less pronounced effects of these drugs on the cognitively better off and more pronounced effects on the cognitively worse off suggest that they would not increase and might decrease social inequality. Finally, empirical evidence for the potential of drugs to enhance moral sensitivity and moral motivation will be considered and some of the social implications of the effects of the drugs will be explored. If the drugs were effective in making us more responsive to moral reasons for and against

W. Glannon
Department of Philosophy, University of Calgary, Calgary, AB, Canada
e-mail: wglannon@ucalgary.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_93,

beneficial and harmful actions and promoted moral progress, there would remain the question of how to balance the collective interest against individuals' right to choose not to enhance.

Introduction

Neuroenhancement is the improvement of brain and associated mental functions beyond what is considered normal for the general population (Juengst 1998). It includes improving such cognitive capacities as information processing, concentration, and practical decision-making. It also includes improving the cognitive and affective capacities necessary for moral reasoning about actions that benefit or harm others. While there has been speculation on the possibility of enhancing neural and mental functions through genetic modification, this type of intervention is still unproven. In contrast, many people have enhanced mental functions through psychopharmacology for generations, and this practice is becoming more pervasive especially among adolescent and young adult populations. Some studies estimate that approximately 25 % of American secondary school students use psychostimulants such as dextroamphetamine (Adderall), methylphenidate (Ritalin), and the wakefulness-promoting drug modafinil (Provigil) for the nontherapeutic purposes just mentioned (Wilens et al. 2008; Greely et al 2008). Earlier studies estimated that roughly 7 % of US university students used these drugs for the same purposes (McCabe et al 2005). Roughly 5 % of the working population in Germany uses psychotropic drugs to enhance their cognitive functions (Heinz et al. 2012). In 2008, the journal *Nature* published the results of an informal survey of its readers (Maher 2008; Sahakian and Morein-Zamir 2007). One-third of the 1,400 readers who responded said that they had used these drugs for off-label nontherapeutic reasons. The use of these drugs to perform better on exams, write more successful grant applications, or improve work performance is likely to increase.

Some opponents of neuroenhancement claim that it will alter and adversely affect our human nature (Kass 2003; Sandel 2007). Others point out that the risks of chronic use of psychotropic drugs to improve mental functions may outweigh the benefits (Heinz et al. 2012). Still others claim that enhancement will increase social inequality and injustice (Habermas 2003). Proponents argue that human nature is not fixed but evolving and has many imperfections that provide compelling reasons for using drugs to improve our capacity for practical and moral reasoning (Harris 2007). They argue that cognitive enhancement will not necessarily result in a zero-sum world of increased competition but instead will increase productivity and result in significant benefits for both individuals and society (Buchanan 2011a, b). Moreover, they argue that enhancing our limited moral capacities will be especially beneficial because it will promote pro-social behavior and reduce the harm resulting from self-interested behavior. Moral enhancement could help to defuse two particularly pressing threats to the survival of human and nonhuman species: the potential use of weapons of mass destruction and environmental degradation (Persson and Savulescu 2011, 2012).

This chapter will first address the claim that neuroenhancement will have a negative impact on human nature. It will be shown that this claim is largely unfounded and rests on questionable assumptions without adequate supporting arguments. The chapter will then consider the risk of addiction from chronic use of cognition-enhancing drugs. A definitive answer to the question of whether this risk is significant enough to limit or prohibit their use will not be available until a sufficient number of long-term controlled studies of the drugs' effects have been conducted. Based on evidence from studies conducted thus far, more pronounced effects of these drugs on the cognitively worse off and less pronounced effects on the cognitively better off would not increase and might even reduce social inequality. There will be no discussion of mood enhancement in this chapter. There was considerable debate in the 1900s and early 2000s about how selective serotonin reuptake inhibitors (SSRIs) might improve the mood of people not diagnosed with clinical depression. Recent meta-analyses have demonstrated that SSRIs may be no more effective than placebos in mild to moderate depression (Kirsch et al 2008; Fournier et al 2010). This suggests that anecdotal claims of enhanced mood by people who are not clinically depressed may reflect nothing more than a placebo response and not the result of the drugs effect on the brain. Much of the recent debate on enhancement has focused on moral enhancement and how psychopharmacological interventions might increase our capacity for empathy, trust, and cooperation by strengthening our capacity to respond to moral reasons for and against certain actions. The empirical question of whether and to what extent the relevant drugs could have these effects will be addressed, as well as the normative question of whether individuals would be obligated to enhance their moral sensitivity or would retain their right to refuse it.

Altering Human Nature

Opponents of neuroenhancement generally object to it for two reasons: (1) that it will fundamentally change our human nature and undermine its intrinsic value and (2) that it is symptomatic of the hubristic quest for perfection and the failure to accept the fact that much of what we do and achieve is beyond our control. In both respects, presumably neuroenhancement would have only a detrimental effect on our character and make all of us worse off than we would be without it.

The argument from nature assumes that the biological and psychological characteristics that define human nature are essentially good and accordingly should not be tinkered with. Yet the fact that the natural design of our bodies makes us susceptible to a range of physical and mental diseases clearly shows that nature is not always good. Enhancing the regenerative potential of somatic cells and tissues using induced pluripotent stem cells, for example, could benefit us by preventing or delaying the onset or progression of many diseases. In addition, enhancing our cognitive capacity to process large amounts of information and anticipate the future could improve our ability to predict pandemics and develop vaccines that could

prevent them from disabling or killing us. There may be an evolutionary advantage to enhancement. It may enable us to adapt faster to environmental changes that could threaten our survival. If we do not take any risks in trying to improve our physical and mental traits, then we may be more vulnerable to natural forces beyond our control. Altering our given biological nature can make us better off by improving our adaptability; not altering it can make us worse off in this regard. Neurocognitive enhancement would not undermine our capacity to make normative judgments about the good and bad. This capacity is not an intrinsic property of human nature as such but of an evolved cognitive and affective capacity to recognize and respond to reasons for or against actions that affect people in positive and negative ways. The fact that we can make judgments about the presumed goodness of our nature suggests that we have a conception of goodness that is independent of our nature, or at least that it is not entirely a function of it (Buchanan 2009, 2011a, b, p. 73). Insofar as enhancing our cognitive capacities made us more rational and less biased in our value judgments, it would promote impartiality and fairness in our actions and assessment of human behavior.

Claims by Leon Kass and others that human nature is good and that altering it would be bad begs the question by assuming what needs to be proven (Kass 2003). Allen Buchanan articulates the main problem with the naturalist position: “If a biomedical intervention had the unwanted consequence of destroying the capacity for judging goodness, then it would follow trivially that *this* alteration of human nature undercut our capacity for judging goodness. But it would not follow that *any* alteration of our nature would undercut that capacity” (2009, p. 150). Indeed, as noted, altering our nature in the respects that have been described could improve it and thereby benefit us. Those who appeal to human nature to criticize enhancement ignore the fact that there are both salutary and deleterious aspects to our biological and psychological design.

Michael Sandel claims that the desire to enhance our cognitive and physical capacities is driven by the desire for perfection and absolute control of our lives (Sandel 2007). We should welcome rather than try to alter our natural cognitive and physical traits, imperfect as they are. It is with reference to this that Sandel says that we should appreciate the “giftedness” of life and cultivate an attitude of openness to the “unbidden.” While it is important to distinguish between the gifted and the merely given aspects of human nature (Hauskeller 2011), Sandel is particularly concerned about the harmful effects that enhancement would have on our character.

But the desire to *improve* our capacities and character does not imply a desire to *perfect* them. Nor does it reflect a hubristic desire to completely control our lives. To cite Buchanan again: “Even in a world of pervasive and powerful biomedical enhancements, we’d still have plenty of opportunities for appreciating that many of the good things in our lives are not our accomplishments, not subject to our wills” (2011b, p. 134). Enhancement could have unintended bad effects on character. Yet it is not enhancement as such but its effects on our attitudes that would determine its value or disvalue. One cannot assume without evidence or argument that enhancement would have only adverse effects on us. It would be difficult to find anything morally objectionable about improving our character, especially if it is constituted

primarily by self-interest that neglects the needs, interests, and rights of others. Also, Sandel fails to adequately appreciate inequalities in people's natural cognitive abilities. Improving these abilities among those who are cognitively worse off could make for a more egalitarian and just society. Claims that enhancement would undermine human nature are vague and unsupported. Buchanan rightly claims that "normative essentialist appeals to human nature ought to be rejected" and that "the enhancement debate is more fruitfully pursued in other terms" (2009, p. 150). Let us now consider some of these other terms.

Cognitive Neuroenhancement

For those with normal cognitive functions, drugs such as methylphenidate, dextroamphetamine, and modafinil purportedly improve attention, concentration, and other cognitive functions associated with working memory. Yet short-term studies have shown that people with a higher cognitive baseline on an absolute scale tend to benefit less from these drugs, while those with a lower baseline tend to benefit more (Farah et al. 2004; de Jongh et al. 2008; Repantis et al. 2010). The drugs appear to display an inverted dose–response curve, with lower concentrations improving functions and higher concentrations not affecting or impairing functions. Dopamine is the main neurotransmitter targeted by these drugs. Children with attention-deficit/hyperactivity disorder (ADHD) tend to do better academically when taking methylphenidate and other stimulant medications than children with the same disorder who do not take them (Scheffler et al. 2009). These medications can improve cognitive control over their thought and behavior (Hyman 2012). But higher-than-normal concentrations of dopamine do not always improve and may interfere with some cognitive functions. Some studies have shown that methylphenidate improves cognitive functions on novel tasks but impairs functions on learned tasks (Elliott et al. 1997; de Jongh et al. 2008). So there may be cognitive trade-offs in using these drugs. Provided that the doses are low to moderate, however, there is no evidence to suggest that occasional use of these drugs in preparing for or taking an exam or writing a grant application would have any untoward effects on those who use them.

But chronic use has the potential to cause addiction and other forms of maladaptive and pathological behavior. The dopaminergic pathways targeted by these drugs play a key role in the brain's reward system. Hyperactivation of this system by consistently high concentrations of dopamine may impair inhibitory mechanisms on impulsive behavior in the prefrontal cortex and impair behavioral flexibility and voluntary control of one's actions (Volkow et al. 2009). Because some cognitive functions may be improved at the cost of others being impaired, we might have to qualify the use of "psychopharmacological enhancement." Together with the risk of addiction, chronic use of drugs like methylphenidate and dextroamphetamine would be expected to have the side effects associated with the general class of amphetamines. These would include, but not be limited to, hypertension, increased risk of stroke and myocardial infarction, insomnia, and, in some cases,

psychosis. Some people may use the beta-adrenergic receptor antagonist propranolol for cognitive enhancement. Unlike methylphenidate, propranolol does not directly affect brain mechanisms in the prefrontal cortex regulating executive functions and the capacity for attention and focus. Instead, it attenuates the autonomic response to stress by preventing or reducing a rapid heartbeat, sweating, and other bodily manifestations of performing under pressure as a concert musician or public speaker. The effects of propranolol would prevent these symptoms from interfering with the individual's focus on the cognitive tasks at hand and thereby enhance the person's capacity to execute these tasks. Yet this drug can dampen noradrenergic mechanisms in the amygdala regulating the capacity to experience fear (LeDoux 1996; Stahl 2006). This could interfere with a natural neural and psychological response to fear-inducing stimuli and put an individual at risk of harm. Specifically, by diminishing this response, the drug could cause a pedestrian to be less cautious while navigating streets in urban settings with heavy traffic volumes. Also, by taking some of the edge off of the anxiety experienced by a musician or public speaker when performing in a concert or giving a speech, the drug could attenuate the emotional valence associated with these social interactions and detract from the meaning and positive feeling they ordinarily entail.

Again, large-scale, long-term prospective studies need to be conducted to accurately assess the benefits and risks of cognition-enhancing drugs. Until this occurs, as a matter of autonomy, individuals should have the right to use these drugs if they so choose. To be sure, marketing strategies by the pharmaceutical industry touting the "benefits" of the drugs could distort the information available to those considering using them. But it would be paternalistic to suggest that these strategies would prevent most people from being suitably informed of the potential benefits and risks of enhancement. They could still freely choose to enhance or abstain from enhancing their cognitive capacities. The influence of the industry would not be sufficient grounds for prohibiting people from using the drugs. At the same time, though, those who used the drugs should take responsibility for any untoward consequences and be accountable for any social costs associated with them. Specifically, in principle they should pay for any medical care necessary to treat conditions resulting from using the drugs.

The risk of these drugs should be enough to prohibit physicians from prescribing them to individuals with normal cognitive functions. Consistent with the professional obligation to benefit and not harm patients by preventing or treating disease, there would be no justification for prescribing a drug to enhance cognitive functions if there were no disease to prevent or treat and there were any risk involved in taking it (Beauchamp and Childress 2008, Chaps. 5 and 6). Prohibiting this practice could also be justified on the grounds that prescribing cognition-enhancing drugs to healthy people with normal cognitive functions would not be an efficient use of limited resources in a publicly funded health care system (Merkel et al 2007, p. 37; Forlini et al. 2012; cf. Synofzik 2009). Of course, individuals may and indeed do often acquire cognition-enhancing drugs from acquaintances or the Internet. But professional medical ethics should prevent physicians from being an additional source of the drugs.

Some might claim that there are no substantial differences between the cognition-enhancing effects of psychostimulants and those of substances in certain foods and drinks, such as caffeine. Yet Andreas Heinz and coauthors point out that "there is a substantial difference between dopamine release caused by a particular experience of food without addiction potential and one resulting from psychotropic substances used for the purpose of neuroenhancement with respect to the extent and the habituation of dopamine release" (Heinz et al. 2012, p. 373). They add: "Dopamine concentrations, which are triggered by food, sex, and human communication, increase by approximately 50% to 100%, while drugs that directly affect dopamine transporter function (e.g., amphetamine, cocaine, and the psychostimulants methylphenidate and modafinil) induce dopamine releases ranging from 175% to 1000%. As a result, subjects learn to crave for the 'more effective' drugs of misuse and lose interest in nondrug-associated stimuli." (p. 373). Repeated use of these drugs can cause dysregulation of dopaminergic mechanisms in the brain. In addition, caffeine appears to stimulate dopamine release in the prefrontal cortex but not also in the ventral striatum, which is a component of the reward system. The effects of psychotropic drugs in the brain are more direct and pronounced than those of substances in our daily food and drink. If chronic use of these drugs inclined some people to addictive behavior, then we could question whether cognitive enhancement had a favorable benefit-risk ratio.

Research into the addictive potential of cognitive enhancement may pose an ethical problem regarding the determination of acceptable risk and the safety of the drugs. Participants in long-term studies of the drugs would be healthy subjects with no history of addiction. Those in the experimental arm of a clinical trial receiving regular doses of a psychostimulant and experiencing heightened dopaminergic effects in their brains' reward system would be exposed to a potentially addictive drug. Proving that chronic use of a drug was in fact addictive could mean causing previously healthy individuals to become addicted! Heinz et al. say that "to date no single variable has been identified that reliably predicts addiction to neuroenhancers. To identify such predictive factors, a substantial number of subjects would have to be exposed to potentially addictive drugs in prospective studies. Since these healthy volunteers do not have a disease, the risk-benefit ratio for such exposures remains rather unfavorable." (p. 374). Subjects should retain the right to participate in these studies if they are capable of giving informed consent to participate in them, and this would include being informed of any risks. But some may cite researchers' duty of nonmaleficence and question whether exposing healthy subjects to potentially addictive behavior would be consistent with this duty. With careful design and monitoring of effects in subjects, studies that quantified the risks associated with chronic use of cognition-enhancing drugs would be both empirically and ethically justified. Still, this research would require resources from public health systems. Because health resources are scarce, many would argue that research into the effects of cognitive enhancement should receive lower funding priority than research into developing preventive and therapeutic treatments for neuropsychiatric and other diseases. There is a more urgent need to generate scientific knowledge of the causes and possible treatments for these diseases than to generate knowledge about the effects of trying to improve normal brain function. Indeed, because it does not involve disease, some might argue

that research into the long-term effects of cognitive enhancement should not be funded at all.

Some opponents of cognitive neuroenhancement claim that it would exacerbate unjust social inequalities if it were practiced on a broad scale. Although people decide to take cognition-enhancing drugs for many reasons, one reason is to give them a competitive edge over others in obtaining such positional goods as an elite education or lucrative jobs. It may seem unfair that some who are already cognitively better off than others could improve their cognitive capacities. Enhancement already occurs when parents arrange for private tutors for their children or send them to private schools. It also occurs when athletes employ private trainers. So we should not focus entirely on drug-induced cognitive enhancement. But these examples of enhancement illustrate income inequalities between the better off and worse off, which suggests that it is unfair to those who lack the financial means to have these same opportunities. Nevertheless, if the positive effects of cognition-enhancing drugs are more pronounced among the cognitively worse off and less pronounced among the cognitively better off, and if the drugs were accessible to all, then wider use of the drugs by more people would not likely increase inequality and might do more to approximate a level social playing field.

If scientific studies and not just anecdotal evidence demonstrated that the drugs had positive effects on executive functions, then some students, university teachers, and other workers might feel coerced into taking the drugs just to keep up with their cognitively enhanced peers. Employers might pressure employees to take the drugs in order to work more efficiently and productively. This already occurs to some extent in the military. Some personnel are encouraged or required by their superiors to take modafinil or other stimulants in order to perform certain tasks over long periods without the need for sleep. Some workers might decline to take the drugs out of concern about potential deleterious effects on their bodies and brains. This could put them at a competitive disadvantage compared with those who took the drugs. Whether these concerns were warranted would depend on the actual effects on neural and cognitive functions. On the basis of meta-analyses of the effects of methylphenidate and modafinil in healthy individuals, psychopharmacologist Reinoud de Jongh and coauthors state: “High expectations regarding the effects of enhancement drugs are not warranted. However, societal pressure may occur with respect to drugs that are ineffective or only slightly effective simply because people believe these drugs improve performance, as the illicit use of methylphenidate and modafinil shows. Creating realistic expectations appears to be very hard to accomplish” (de Jongh et al 2008, pp. 771–772; see also Repantis et al. 2010; and Lucke et al. 2011).

Coercion can result from unreasonable expectations impose by some on others, forcing them to accept situations that are worse for them than any alternatives. But if everyone were suitably informed of the generally modest effects of psychostimulants on cognition, then expectations about these effects would adjust accordingly. Consequently, the incidence of real or perceived coercion or unfairness regarding the use of these drugs would diminish. Moreover, drugs that enhanced the relevant cognitive capacities would not alone determine a better or

worse outcome. Individual effort in exercising these capacities and time management would also be necessary. The student would still have to write and submit the term paper on time, and the researcher would still have to write and submit the grant application by the deadline. Another factor to consider is that enhancing cognitive capacities by itself would not necessarily make our lives better on the whole, since well-being is more than a function of these capacities (Tannsjo 2009).

In 2008, a group of neuroscientists and ethicists formulated a set of recommendations based on the presumption that mentally competent adults should be able to engage in cognitive enhancement (Greely et al 2008). These recommendations included an evidence-based approach to the evaluation of risks and benefits and enforceable policies regarding the use of cognition-enhancing drugs to support fairness, protect individuals from coercion, and minimize enhancement-related socioeconomic disparities. Even if the risks were proven to be significant, it would not be clear whether this would warrant a limitation or prohibition on enhancement and whether competent, autonomous individuals could lose their right to enhance. This would come with the proviso that they would be responsible for any social costs resulting from enhancement.

Moral Enhancement

A potentially more beneficial intervention in the brain and mind would be psychopharmacological enhancement of our moral disposition and sensitivity in recognizing and responding to the rights, needs, and interests of others. This would involve not only improving the cognitive capacity to respond to reasons but also the affective capacity for social emotions such as shame, regret, and remorse. It could improve the cognitive capacity to imagine counterfactual situations, making us more cautious in deliberating about which actions to perform. It could also improve the ability to foresee the probable consequences of our actions. Increasing our capacity for extended empathy, trust, and cooperation could result in moral progress by promoting behavior that would reduce the incidence of humans harming other humans and result in greater well-being for all (Douglas 2008; Spence 2008; Persson and Savulescu 2011, 2012; Harris 2012). Among other things, enhancing our moral capacity could be an antidote to the excessive competition and pervasive cheating in academia and sport that have been associated with enhancing cognitive functions.

As with cognitive neuroenhancement, genetic and pharmacological modification of our moral thought and behavior have been suggested as different possible ways of producing the desired effects. Genetic modification of the neural processes mediating our moral sensitivity would not likely be feasible. Our neural and mental characteristics are not just a function of our genes but also of epigenetic factors associated with our bodies and the environment. Genotype does not determine phenotype. So even if genes or gene products associated with behavior could be delivered safely into the relevant regions of the brain, this would not ensure that the individual would display enhanced moral behavior. In light of these and other

considerations, Buchanan says: “My hunch is that widespread genetic enhancements of humans will not occur for a long time, if ever” (2011b, p. 100). Psychotropic drugs targeting particular neurotransmitters would have more direct effects in the neural networks mediating moral capacity, and these effects could be monitored in real time with functional neuroimaging.

Good bioethics relies on some degree of speculation in anticipating ethical issues arising from biotechnology. On the issue of moral enhancement, however, we need to be careful not to put the conceivability-driven ethical cart before the empirical horse. In their recent book, *Unfit for the Future*, Ingmar Persson and Julian Savulescu claim that it is possible to change our moral attitudes to avert threats from weapons of mass destruction and environmental degradation to our well-being and survival (Persson and Savulescu 2012). They also claim that it is conceivable that a pharmaceutical could cause the chemical changes in our brains underlying these attitude changes. Given the urgency of the threats they describe, they say that if it were feasible, effective moral bio-enhancement would be the most important kind of biomedical enhancement. Few would disagree in principle with the idea that we should enhance our capacity for moral reasoning and behavior to reduce or prevent the harm that we cause to present and future generations. But one crucial question is this: How would psychopharmacology induce the necessary neural and attitudinal changes to meet these challenges? Conceivability alone does not provide an adequate framework within which to discuss the ethical issues. Empirical data on the neural targets and psychological effects of selected drugs is needed.

Some studies suggest that increasing concentrations of serotonin in the prefrontal cortex could induce or increase harm aversion and make us more likely to respond to moral reasons against actions that adversely affect others. The SSRI citalopram (Celexa) has been shown to enhance social cognition and promote pro-social behavior by increasing the capacity for harm aversion in healthy subjects (Crockett et al 2010; but see also Chan and Harris 2011). Yet the fear and impaired empathy associated with harmful behavior are mediated by subcortical more so than by cortical brain regions, where serotonin is particularly abundant. The cognitive and affective capacities necessary for moral reasoning are mediated by distributed neural networks in interacting cortical and subcortical regions. The amygdala and anterior cingulate cortex are particularly relevant to fear and empathy. It is doubtful that the effects of a drug targeting serotonin concentrations in the prefrontal cortex would project to these other regions and influence other neurotransmitters modulating fearful and empathic responses.

A drug might produce these modulating effects by increasing concentrations of oxytocin in the brain. There is emerging evidence that this neuropeptide plays a critical role in social cognition (Ross and Young 2009). A drug administered intranasally that increased oxytocin in the brain could promote pro-social behavior and enable us to move beyond self-interest and group bias. Oxytocin interacts with the hypothalamic-pituitary-adrenal axis to inhibit activity in the amygdala mediating the fear response. By diminishing this response, it is possible that such a drug would increase the cognitive-affective capacity for concern, trust, and cooperation. Another neuropeptide, vasopressin, appears to function in the same way. Yet the salutary

effects of drugs aimed at increasing levels of these neuropeptides cannot be assumed. For example, some studies suggest that social context can influence the effects of oxytocin on people's behavior. This may involve how one perceives others, which can be a function of the social group to which they belong. The drug may strengthen the experience of bonding with others in groups with which one identifies and contribute to the perception of those outside these groups as competitors or threats. In some cases, social factors may cause oxytocin to promote mistrust and antisocial rather than pro-social behavior (Bartz et al. 2011). There are questions about the safety and efficacy of chronic administration of oxytocin and vasopressin, which can only be answered after more long-term controlled clinical trials have been conducted.

Given the extent of actual and potential harm resulting from the failure to recognize and respond appropriately to the rights, needs, and interests of others, there are compelling reasons for developing and implementing safe and effective moral enhancement. But at least four issues would have to be addressed and resolved before doing this.

First, extensive research would have to be conducted to identify a drug or drugs that would safely and effectively enhance our moral sensitivity. This would be costly, and it raises the same question as the earlier question regarding research on cognition-enhancing drugs: Should research into psychopharmacological agents designed to enhance our moral behavior be treated on a par with research designed to prevent or find treatments for neuropsychiatric and other diseases? Some would argue that the harm resulting from actions symptomatic of our limited moral compass is as significant as, if not more so than, the harm from disease. But the immediacy and salience of the latter appear to give it more medical and moral weight. What complicates drawing accurate and fair comparisons between them is that these are distinct types of research with distinct aims. Professional and public debate will be necessary to decide whether priority in resource allocation should go to research aimed at ameliorating flaws in our bodies and brains or to research aimed at ameliorating flaws in our character.

Second, it is doubtful that pharmacological modification of our cognitive and affective capacities alone would make us more responsive to moral reasons when acting. Education thus far has failed to do this, as Persson and Savulescu point out. Yet the right type of education could complement psychopharmacology in making us more responsive to the interests of those who exist now and those who will exist in the future. The social environment also influences brain function and provides cues that prompt us to act in different ways. Moral enhancement would thus require a full complement of education, environmental modification, and psychopharmacological intervention.

Third, just because a drug might enhance our capacity for empathy, trust, care, and concern for others does not mean that we would exercise this capacity. We may fail to put the necessary effort into translating receptivity to moral reasons into morally defensible actions. Factors in the social or cultural environment often motivate us to act. Yet all too often these factors incline us to self-interest and group bias. It is unclear whether or to what extent the appropriate drugs could weaken the influence of these factors on our behavior. Indeed, as I have noted, drugs such as oxytocin may

exacerbate this bias. These considerations reinforce the earlier point that enhancement of our moral attitudes would depend on a tripartite model of education combined with modification of the social environment and our cognitive and affective capacities. Alluding to this idea, Buchanan says that “a biomedical intervention might be one aspect of a multifaceted effort to extend concern and respect for all human beings, not just those who are like us” (2011b, p. 170).

Fourth, some proponents of moral enhancement insist that it must be a collective enterprise in which people are not merely permitted but obligated to enhance their cognitive and affective capacities (Harris 2007, 2011, 2012; Persson and Savulescu 2012). Reducing or averting harm from the most serious threats to our well-being and survival is a collective action problem. As such, this and related goals can only be achieved if a sufficient number of people engage in moral enhancement. Given the choice, though, many people would decide not to enhance because they lacked the motivation to strengthen their moral disposition. If a significant number of people opted out of enhancement, then it could preclude the collective effect from being achieved. There would also have to be agreement on which attitudes or virtues should be enhanced and which vices should be diminished. This would presuppose a common understanding of the concept and goals of moral enhancement by the general public. It cannot be predicted or assumed that this critical agreement could be achieved. In addition, there would have to be agreement on which drugs would be most likely to safely and effectively enhance the disposition to be moral. Among other factors, this would assume that pharmaceutical companies producing the drugs would be acting in good faith.

Assuming that there could be public consensus on these issues, making moral enhancement mandatory rather than voluntary would involve a different and more controversial set of problems. It could mean that individual liberty in choosing whether or not to enhance could be overridden by collective interest. This would be a violation of the presumptive right to noninterference in one’s body, including the brain and mind. Given the magnitude of the actual harm from people failing to respect the rights and interests of others, the potential harm from the use of weapons of mass destruction and environmental degradation, and assuming that psychopharmacology could improve our moral behavior, some restrictions on individual liberty in choosing not to enhance might be justified. The core concern is not the benefit-risk ratio in using psychoactive drugs but more fundamentally the right to refuse what may be a beneficial intervention for individuals and society. Even if some restrictions on individual liberty for collective benefit were intuitively acceptable, we would still have to identify, collectively agree on, and morally and legally justify what these restrictions would be.

Conclusion

I have considered the respects in which our cognitive and moral capacities can be enhanced through psychopharmacology. After examining the empirical issue of how drugs might enhance these capacities, I considered some of the ethical reasons

for and against enhancement. Claims that enhancement would result in a deleterious alteration of human nature, and that it would involve a hubristic pursuit of perfection, are unfounded. Instead, what matters morally is whether drugs altering our neural and mental functions would have unintended bad consequences on these functions. I noted evidence pointing to the potential of cognition-enhancing drugs to cause addiction. Citing the recommendations of an eminent group of neuroscientists and ethicists, I also noted that questions about the safety and efficacy of these drugs could only be answered after a significant number of long-term studies had been conducted. There are ethical questions about the research itself, since it would involve healthy subjects who might be exposed to a risk of addictive behavior by participating in these studies. Given limited health resources, there is the additional question of whether funding research into improving normal neural and mental capacities could be justified.

Moral enhancement has the potential to be significantly more beneficial than cognitive enhancement because it could have positive effects on large numbers of people in reducing harm, increasing well-being, and averting threats to the survival of human and nonhuman species. By improving our moral sensitivity, it could promote moral progress. Normative questions about whether we should enhance our capacity for empathy, care, and trust would depend on empirical evidence showing that certain drugs could produce these changes in our attitudes toward others. Some have argued that the collective effects of moral enhancement that are necessary to avert the gravest threats to our survival can only be achieved by requiring individuals to engage in it as part of a collective enterprise. This may involve some restrictions on individual liberty and the presumptive right to refuse an intervention in one's brain and mind. Not everyone would agree that the interests of the collective could justify limiting the rights of individuals to refuse enhancement, even if it had great promise in making all of us better off. Given the prospect of moral enhancement, the ethical challenge would be to strike the right balance between respecting the rights and interests of individuals and those of the human community. Only by striking this balance can moral progress be made.

Acknowledgments The author is grateful to Bert Gordijn for helpful comments on an earlier version of this chapter.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Human Brain Research and Ethics
▶ Impact of Brain Interventions on Personal Identity
▶ Moral Cognition: Introduction
▶ Neuroenhancement
▶ Research in Neuroenhancement

► Sensory Enhancement
► Smart Drugs: Ethical Issues
► The Morality of Moral Neuroenhancement
► What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Bartz, J., et al. (2011). Social effects of oxytocin in humans: Context and person matter. *Trends in Cognitive Sciences, 15*, 301–309.

Beauchamp, T., & Childress, J. (2008). *Principles of biomedical ethics* (6th ed.). New York: Oxford University Press.

Buchanan, A. (2009). Human nature and enhancement. *Bioethics, 23*, 141–150.

Buchanan, A. (2011a). *Beyond humanity? The ethics of biomedical enhancement*. New York: Oxford University Press.

Buchanan, A. (2011b). *Better than human: The promise and perils of enhancing ourselves*. New York: Oxford University Press.

Chan, S., & Harris, J. (2011). Moral enhancement and pro-social behavior. *Journal of Medical Ethics, 37*, 130–131.

Crockett, M., et al. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107*, 17433–17438.

de Jongh, R., et al. (2008). Botox for the brain: Enhancement of cognition, mood and pro-social behavior and blunting of unwanted memories. *Neuroscience and Biobehavioral Reviews, 32*, 760–776.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*, 228–245.

Elliott, R., et al. (1997). Effects of methylphenidate on spatial working memory and planning in healthy young adults. *Psychopharmacology, 131*, 196–206.

Farah, M., et al. (2004). Neurocognitive enhancement: What can we do and what should we do? *Nature Reviews Neuroscience, 5*, 421–425.

Forlini, C., et al. (2012). Should physicians prescribe cognitive enhancers to healthy individuals? *Canadian Medical Association Journal*. doi:10.1503/cmaj.121508.

Fournier, J., et al. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medical Association, 303*, 47–53.

Greely, H., et al. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456*, 702–705.

Habermas, J. (2003). *The future of human nature*. Cambridge, UK: Polity Press.

Harris, J. (2007). *Enhancing evolution: The ethical case for making better people*. Princeton: Princeton University Press.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*, 102–111.

Harris, J. (2012). Moral progress and moral enhancement. *Bioethics*. doi:10.1111/j.1467-8519. 2012. 01965.x.

Hauskeller, M. (2011). Human enhancement and the giftedness of life. *Philosophical Papers, 40*, 55–79.

Heinz, A., et al. (2012). Cognitive neuroenhancement: False assumptions in the ethical debate. *Journal of Medical Ethics, 38*, 372–375.

Hyman, S. (2012). Might stimulant drugs support moral agency in ADHD children? *Journal of Medical Ethics*. doi:10.1136/medethics-2012-100846.

Juengst, E. (1998). What does enhancement mean? In E. Parens (Ed.), *Enhancing human traits: Ethical and social implications* (pp. 29–47). Washington, DC: Georgetown University Press.

Kass, L. (2003). *Beyond therapy: Biotechnology and the pursuit of happiness*. New York: Harper Collins.

Kirsch, I., et al. (2008). Initial severity and antidepressant benefits: A met-analysis of data submitted to the Food and Drug Administration. *PLoS Medicine, 5*(1), e45.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

Lucke, J., et al. (2011). Deflating the neuroenhancement bubble. *AJOB Neuroscience, 2*(4), 38–43.

Maher, B. (2008). Poll results: Look who's doping. *Nature, 452*, 674–675.

McCabe, S., et al. (2005). Non-medical use of prescription stimulants among US college students: Prevalence and correlates from a national survey. *Addiction, 100*, 96–106.

Merkel, R., et al. (2007). *Intervening in the brain: Changing psyche and society*. Berlin: Springer.

Persson, I., & Savulescu, J. (2011). Getting moral enhancement right: The desirability of moral bioenhancement. *Bioethics*. doi:10.1111/j.1467-8519.2011.01907.x.

Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.

Repantis, D., et al. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*, 187–206.

Ross, H., & Young, L. (2009). Oxytocin and the neural mechanisms regulating social cognition and affiliative behavior. *Frontiers in Neuroendocrinology, 30*, 534–547.

Sahakian, B., & Morein-Zamir, S. (2007). Professor's little helper. *Nature, 450*, 1157–1159.

Sandel, M. (2007). *The case against perfection: Ethics in the Age of genetic engineering*. Cambridge, MA: Harvard University Press.

Scheffler, R., et al. (2009). Positive association between attention-deficit/hyperactivity disorder medication use and academic achievement during elementary school. *Pediatrics, 123*, 1273–1279.

Spence, S. (2008). Can pharmacology help enhance human morality? *The British Journal of Psychiatry, 193*, 179–180.

Stahl, S. (2006). *Essential psychopharmacology* (2nd ed.). New York: Cambridge University Press.

Synofzik, M. (2009). Ethically justified, clinically applicable criteria for physician decision-making in psychopharmacological enhancement. *Neuroethics, 2*(2), 89–102.

Tannsjo, T. (2009). Ought we to enhance our cognitive capacities? *Bioethics, 23*, 421–432.

Volkow, N., et al. (2009). Effects of modafinil on dopamine and dopamine transporters in the male human brain. *Journal of the American Medical Association, 301*, 1148–1154.

Wilens, T., et al. (2008). Misuse and diversion of stimulants prescribed for ADHD: A systematic review of the literature. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 21–31.

Section XVII

Neurolaw

Neurolaw: Introduction

81

Reinhard Merkel

Contents

Abstract

In recent years, a host of normative questions have surfaced in the wake of rapid progress in the neurosciences and have entered the sphere of legal discourse. They continue to confront legal scholars and increasingly also courts with problems that strike many lawyers as rather unfamiliar within their accustomed domain. Some of these problems seem to be entirely new, others appear to shed new light on venerable and unresolved matters of the law and its philosophical foundations. In a rough and ready differentiation, one can distinguish three perspectives from which to assess these developments. Neuroscience provides (1) new insights (into old metaphysical problems of mind and brain), (2) new ways to peer into, and (3) new ways to intervene in the brain (and hence, perhaps, the mind). The realm of (1) may be fittingly exemplified by the recent outpouring of neuroscientific contributions to the age-old philosophical question of freedom of the will and the attached problem of legal responsibility; (2) refers to new methods of neuroimaging, visualising structural as well as functional features of the living (and working) brain; and (3) points, most notably, to new ways of enhancing mental capacities of healthy people by altering neural circuits in their brains. – Each of the entries in the present section contributes to (at least) one of these spheres of new questions.

Christoph Bublitz and Martin Dresler deal with legal problems posed by possible manipulations of memory. Memory is not a single entity but a complex phenomenon encompassing several mental capacities and various subjective states of

R. Merkel
Faculty of Law, University of Hamburg, Hamburg, Germany
e-mail: reinhard.merkel@jura.uni-hamburg.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_81,

mind. Many of its puzzling features, therefore, belong to the perennial problem of the relation of mind and brain, though up to now it has been treated rather stepmotherly by professional philosophers of mind. By contrast, empirical, particularly neuroscientific research on memory has remarkably blossomed within the last decades. It has produced a host of results that more or less directly affect concerns of the law.

Legal systems, especially in their procedural parts, rely on accurate memory functions in a number of ways, e.g., on memories of witnesses or of victims of crimes. Over and above this concern for its own functioning, however, the law may also have to engage with memory as a constitutive feature of personhood, indispensable, e.g., for the autobiographical "self" of persons. For this part of memory has become vulnerable in recent years, and hence a legitimate objective for an individual claim to legal protection. As Bublitz/Dresler point out, neuroscience might be on the brink of the capability of modulating memories in various ways, most notably by erasing them, altering or editing their content, or attenuating their vividness.

The authors start with elucidating the phenomenon of memory as well as the corresponding concept by a set of empirical and conceptual distinctions. "Memory" comprises a bundle of functionally different (though related) capacities and, accordingly, of referring concepts – e.g., "declarative" (semantic, episodic, autobiographic) as opposed to "non-declarative" memory (internalized in certain physical skills acquired over one's life course). Phases of acquiring, storing and recalling memory are distinguished. The underlying neural processes do not take place in neatly separable areas of the brain but rely on complex neural networks and their multifarious interactions. "Forgetting" as the factual and conceptual counterpart of memory is also analysed in its accordant manifoldness. It ought not to be conceived of simply as a mental deficit. On the contrary, to some extent and within normal limits it fulfills important purposes indispensable for our psychic well-being. The authors demonstrate that, rather than being opposites, remembering and forgetting are complementary processes which should both be valued in their own right.

Neuroscientific memory research is still in its infancy; many empirical questions are not well understood yet. However, certain forms of memory have already become accessible to altering interventions from outside, in certain contexts even without the affected person's awareness. Memory alterations might occur in the form of enhancing, diminishing (even erasing) or changing the content of memory. All of them obviously raise challenges for ethics and law.

Bublitz/Dresler begin with an ethical analysis. Part of it gains acuteness by being contrasted with the report "Beyond Therapy," published in 2003 by the U.S. "President's Council on Bioethics." This report postulates an almost sacrosanct duty of individuals as well as societies to "remember fitly and truly" their own past and to not employ neurotechnical means to tinker with memories. Of course, interests of individuals in developing and maintaining a coherent self *prima facie* also speak in favour of preserving accurate reminiscences of one's own past. On the other hand, persons may have experienced severely traumatic events that threaten

their present and future well-being, and hence have legitimate interests in numbing or erasing the corresponding stressful memories.

Bublitz/Dresler reject the President's Council's strong demand and make normative room, as it were, for a moral right to intentionally amend one's autobiographical memory in order to overcome traumatic experiences casting shadows on one's future flourishing. They defend their position in light of moral demands concerning identity, personality and *prima-facie* obligations of persons to promote their mental development. They do, however, concede various dangers associated with unwise memory alterations, but deny a thoroughgoing moral obligation to leave the natural interplay between one's autobiographical remembering and forgetting untouched. There can be good moral reasons to artificially alter one's memory.

An analysis of questions of law follows. The authors underscore the law's primary task, which is not to warrant "good acting" of its subordinates but rather to maximize reciprocal individual liberty in a peaceful society. Hence, the realm of legal freedom is not coextensive with that of morally approved actions. The advent of methods of intervening in the brain and thus the mind gave rise to heretofore unknown problems of safeguarding the mental sphere. Some mental liberties, such as freedom of conscience and thought, have long been placed beyond governmental control by international covenants or national constitutions. Memory, however, is not among them. States have always taken an interest in acquiring knowledge of certain past events by interrogating witnesses. This evinces a public interest in having memories of people preserved within their natural limits. Individuals, of course, also have strong interests in not becoming targets of external memory interferences that they did not consent to. On the other hand, they might also wish to not be legally deprived of their liberty to give such consent when they want to alter their own memories.

This gives rise to a host of questions of legal theory concerning rights and duties (and their respective limits) of states and of individuals. Bublitz/Dresler analyse such questions with regard to interventions in personal (and, in part, collective) memories. Their starting point is Adam Kolber's proposal of a novel right to "freedom of memory." Decidedly sympathizing with such a right, the authors survey its potential scope which would presumably encompass a right to remember, an accompanying legal protection against memory erasure and (arguably) also a right to enhance one's memorizing capacities. But it might well extend beyond that and also protect a right to forget which deserves legal protection as well. Besides a right to have one's own memories erased at will, it presumably covers a right to not being reminded of something one wishes to forget or has, albeit artificially, already forgotten. This latter right, however, must be balanced against various liberties of others, such as the right of free speech, even if making use of such liberties reminds others of things they want to (and have a right to) forget.

The authors discuss a few more possible limits to that right, some of them originating from public interests which in certain contexts do prevail over individual liberty rights whereas in others they do not. Finally, Bublitz/Dresler turn to the question whether legal responsibility might potentially be affected by memory

modulations. A legal duty of wrongdoers to memorise their deeds in order to preserve their bad conscience is rejected (though an analogous moral duty may exist). Hence, a corresponding right of the state to interfere with a perpetrator's attempt to artificially rid themselves of their guilty memories is denied. In their concluding remarks, the authors underscore the future task of legislators, courts and legal scholars to address problems of memory manipulations much more thoroughly than has been done so far.

In his second contribution, which he authors alone, Christoph Bublitz approaches the fundamental question of cognitive liberty. His considerations are premised on the conviction that legislators around the globe face the inevitable challenge of regulating novel neurotechnologies. This holds, on the one hand, for private use of such technologies, but also for various purposes of the state that may touch upon much more sensitive issues, such as coercive mind-interventions supposedly in the public interest. On that occasion, existing legislation may come under critical review.

The author begins his exploration by pointing out some basic legal principles that pertain to a possible right to cognitive liberty. In liberal democracies, the powers of governments are limited, most importantly by constitutional and human rights law. Interferences with rights guaranteed by constitutional and human rights law have to be justified by prevailing interests. Some constitutional guarantees cannot be interfered with at all. Neuroethical proposals of whatever kind and associated policy recommendations cannot be transformed into law unless they conform to these legal boundaries.

Thus the law must identify and define the fundamental rights possibly affected by state actions more precisely. In the present context, this holds with respect to cognitive liberty on the one hand and to forms of interferences with other people's minds on the other. Currently, the legal situation is unclear in some important aspects – not only for the trivial reason that national legal systems vary, but also because the legal parameters of conduct affecting one's own or other's minds are not sufficiently explored. Cognitive liberty is one of the rights that (actual or potential) regulations of neurotechnologies or governmental powers to intervene into citizen's minds might interfere with. As Bublitz points out, while references to cognitive liberty are rather common in neuroethics, its legal meaning, scope and limits are only poorly defined. Many national legal systems do not recognize such a right. However, in the major human rights treaties cognitive liberty is firmly entrenched under the name of freedom of thought.

The author explores some of the important normative aspects of the right. In theory, it provides far-ranging protection against many types of unwanted mind-interventions. In practice, however, the right does not play an important, not even a respectable role. As Bublitz makes clear, it is a neglected, perhaps an even largely forgotten human right, more of a vague principle rather than an enforceable legal claim. It has thoroughly failed to be effective in many contexts that would seemingly provide genuine fields of its application, such as coerced psychiatric treatment or national drug policies. Many lawyers or policy makers may not even have heard of it. The task for legal theory is therefore to provide an interpretation of the right

that lives up to its importance, and, more generally, to argue for its recognition in national legal systems.

The author shows that the right has various roots in pre-constitutional days. He explores some of its ancestry and outlines various classic arguments, for instance of Kant and Mill, for its recognition. He argues that its acknowledgment is not merely a contingent political matter, a claim one may favour or reject depending on one's political taste, but rather that it is an inevitable presupposition of any liberal legal system deserving this label. Bublitz forwards some suggestions pertaining to a solid construal of the right and identifies open questions in this context. Furthermore, he develops a specifically legal (as opposed to ethical) view on a contentious neuroethical issue: whether old (traditional) and new (high-tech) means of intervening into other people's minds are intrinsically different or should be judged and treated on par. He shows that there are indeed significant differences from a legal perspective, having to do not so much with the (possibly equal) effects of intervening actions on other's minds, but rather with fundamental liberties of the intervener in making use of the common public space. It is one thing, the author contends, to use one's freedom of speech and provide convincing arguments in order to change other people's opinions even against their previous will, but quite another to bypass their sensory control and implant a new opinion directly in their minds by altering their brains without their prior consent.

Bublitz recurrently tackles one of the salient features of freedom of thought in human rights law, namely that it is protected unconditionally. That is to say that in theory interferences cannot be justified for any reason whatsoever. As a consequence, the law primarily regulates external behavior of citizens, but abstains from ruling their minds. However, this raises an important question: whether such a strict protection of the mind should be maintained. Current legislation demonstrates that many concerns may not be adequately addressed if states do not have any competency at all to restrict the mental freedom of the individual in at least some respects and with regard to at least some pressing issues, such as paternalistic interventions to safe a person from the self-infliction of grave harms.

However, as Bublitz insists, consigning governments the power to alter the minds of their citizens constitutes "a paradigmatic shift from governing conduct to governing minds with far-ranging implications" in various fields of the law and of civil society. It should and hopefully will prompt controversial discussions between ethicists, legal scholars, courts and, last not least, the public.

Reinhard Merkel investigates problems that in all likelihood will be posed to the law by the rapidly developing methods of neuroimaging. This topic belongs to the second sphere of our tripartite scheme of neuroscientific challenges to the law delineated above: that of new ways to peer into the brain and hence arguably the mind. The author confines his analysis to criminal law as his normative, and to functional (as opposed to structural) magnetic resonance imaging (fMRI) as his neuro-technical subject. Though public and civil law will also be affected by developments in neuroimaging, this impact, the author contends, will not be as profound and puzzling as in criminal law. Within the sphere of the criminal law, he focuses on two topics of particularly crucial importance: (1) brain-based lie

detection in criminal proceedings, and (2) "neuroprediction" of future dangerousness of perpetrators, a problem that may become relevant either in the sentencing phase of criminal trials (as in common law jurisdictions) or in special proceedings of preventive detention after a prison term has been served (as in many continental legal systems).

The author begins with a sketch of the intricate basic science involved in fMRI and points out the particular strengths and limits of the method. Then turning to the problem of neuroimaging of deception in court, he analyses problems of validity and reliability of fMRI. The widespread claim that the scientific suitability of the method for any evidentiary purpose in criminal trials is to be denied in toto, is criticised as overreaching. There is, the author holds, no such thing as a uniform standard of validity applying to all probative aims pursued by the parties to the various phases of criminal proceedings. The defendant or his counsel does not have to prove his innocence, whereas proof of the opposite, his guilt, "beyond reasonable doubt" is what falls on the part of the prosecution and finally of jurors and judges. This asymmetry in evidentiary burdens between the two parties has consequences for the question of what is a sufficiently valid scientific method for either of them. Since the benefit of the doubt is on the side of the defendant (but not of the prosecution) and his own aim defined by the modest need of casting only some such doubt on jurors' convictions of his guilt, fMRI might well be an entirely appropriate method for this goal, and providing even weak circumstantial evidence for his veracity of at least some avail to his end is certainly admissible. At the same time, fMRI is as yet unsuitable and hence inadmissible for prosecutorial probative aims. Or so the author contends.

He then considers the widespread skepticism concerning "seductive allures" ascribed to fMRI: its potential to mislead jurors and judges into believing that what they see are a kind of scientifically objective "snapshots" of what happens in a brain during its engagement with cognitive tasks. The author proposes to counter the risk of such crude misunderstandings not by precluding the entire option for defendants (who have a right to provide even weak evidence for their case), but rather by instructing jurors and judges adequately, explaining to them what the images of an fMRI do and what they do not show, and emphasizing that in any individual case their probative value is low. The author concludes his vote for a limited admissibility of fMRI for lie detection with a set of caveats that must be observed in order to avoid any misguided naiveté on the part of the legal personnel.

There is, however, a principled concern which Merkel addresses subsequently. Assuming that in the not too distant future fMRI will become capable of ascertaining thoughts (and lies) with a high degree of validity even from uncooperative suspects, would it then be legitimate to introduce it in criminal proceedings as a regular (if needed, coercive) means of evidence for prosecutorial purposes, too? The author distinguishes between coercing suspects and witnesses. He denies a right of the state to force defendants to undergo fMRI for lie detection, as this would violate their basic right against compelled self-incrimination. Here the following problem arises: defendants' bodies may well be thoroughly examined for evidentiary purposes. Not so their minds, at least not as far as access to their mental

sphere can only be procured by their own compelled cooperation. In other words, testimonial information from suspects must not be obtained coercively, whereas physical information from their bodies may. Now, fMRI for lie detection seems to do both at once: procuring testimonial information (knowledge, memory etc.) by investigating only the part of one's body where such information is stored and from which it can be inferred – the brain.

The author argues that neuroimaging of deception must be classified as obtaining testimonial information. In such cases, the brain is not searched for any of its physical features, but only to gain access to the corresponding mental processes. Hence, compelled fMRI of deception would violate a defendant's fundamental privilege against self-incrimination. Witnesses, on the other hand, do not have such a privilege. They must not refuse (or give untruthful) testimony. Provided that some years from now even coerced fMRI will fulfill standards of validity, it might be argued that witnesses could become legitimate objects of compulsion to testify in a bran scanner. Merkel emphasizes, however, that liberal states should abstain from such a practice. Controlling the access to one's mind is a constitutive feature of personhood. States should not coerce anyone into a brain scanner.

The author concludes with an analysis of the problem of "neuroprediction" destined to assess the future dangerousness of individuals. In such cases, what is at stake is not punishment for a committed crime, but detention for something one has not done (and is only feared to eventually do). Since such treatment of citizens inevitably leads legal orders to the limits of their legitimacy, states are obliged to exhaust all available and suitable methods to clarify the underlying prognoses. Currently, this is regularly done by obtaining the expertise of (usually two) psychiatric expert witnesses. Since such prognoses are contentious and error-prone, Merkel argues for even a duty on the part of the state to adduce further prognostic indicia, provided they meet basic standards of validity. There is scientific evidence that fMRI may already (or in the foreseeable future) meet such standards with regard to two types of potentially dangerous dispositions: pedophilia and psychopathy. Merkel concludes that if this evidence is confirmed, fMRI will indeed be on the brink of entering legal procedures – and vis-à-vis the inevitable problem of assessing future dangerousness for preventive detention, rightly so.

Imogen Goold and Hannah Maslen tackle a question that has arisen with the advent of new ways (real or potential) of intervening in the brain in order to enhance one's mental, most notably cognitive, capacities. Would such enhanced abilities be accompanied by an enhanced legal responsibility when they are (or, at any rate, could be) deployed to fulfill demanding duties that are prone to fault and hence to legal liability? And could there be a legal obligation to take such an enhancer provided it were effective and posed no unreasonable risk? The authors analyse these questions with regard to the English law of negligence only. But since its principles, by and large, are based on fundamental principles of fairness, they can be taken as widely representative for many other legal orders as well.

Goold and Maslen open up their enquiry with a range of quotations from the medical as well as neuroethical literature in which legal requirements to take

enhancers in situations of impending fatigue are called for, most notably for medical doctors. Could there be such an obligation? And could someone be legally liable for a harm that might have been prevented had he or she taken such an enhancer? The analysis centres on three basic questions: First, “the duty question” – is there a duty to take an enhancer (accompanied by a liability for omissions)? Second, “the breach question” – does failing to take the enhancer amount to a breach of the standard of care? And third, “the causation question” – can it be proved that the omission caused the harm?

The authors explore these questions by deploying a hypothetical scenario: a fatigued surgeon after 36 h of work without rest is confronted with an emergency case requiring an immediate surgical intervention that would take about 2 h and that only she is qualified to perform. Though being badly exhausted, she does have a duty to perform the surgery and she would be legally liable for an omission were she to refuse it. She knows that she could take an enhancer. Would the failure to take it constitute a breach of her duty?

The relevant legal standard of care is not determined by the common *factual* practice in hospitals (which at present would certainly not include taking enhancers). Instead, it must be adjusted to the hypothetical measure of a *reasonable* practice, above all with regard to a fair and sensible balancing of risks and benefits. After pondering various arguments, especially that of some non-trivial risks posed by the enhancer to the surgeon’s own health, the authors conclude that currently English courts would be unlikely to hold her responsible for refusing to take it.

Even if that legal hurdle could be overcome, say, with safer enhancers of the future, the subsequent causation question is hard to answer in the affirmative. On a Humean (i.e., counterfactual) account of causality, tacitly adopted by most jurisdictions, an omission is causal for a harmful effect only if the harm would not have occurred “but for” (or without) the omission, i.e., it would have been prevented by the required action. This is problematic here in two regards: First, can it be ascertained “beyond reasonable doubt”? For obvious reasons, that seems hard to do. How does one *prove* what has not, but only could or would have, happened had the course of events been different? And secondly, since every surgical intervention bears risks that may materialize without any fault on the part of the surgeon, the harm could have occurred anyway, even if the enhancer had warranted a faultless performance of the surgery.

The authors then turn to the problems that arise when enhancers cause harm. Could taking an enhancer amount to a breach of duty? (Would it be an *enhancer* then?) Goold and Maslen give a negative answer. Even if causality could be proved “beyond reasonable doubt,” the liability of the surgeon would have to be denied. The harmful effect as a specific consequence of an enhancement (supposedly designed to prevent such effects) appears to remote to be imputed to the person who took it.

After briefly considering and denying an enhanced legal responsibility as a corollary of artificially lifting oneself above the threshold of the regular standard of care, the authors conclude their explorations with thoughts on the divergence

between ethics and law. Though there might be good ethical arguments in favour of a duty to enhance oneself in stressful, error-prone and potentially harmful situations, the law still has good reason to eschew an analogous legal duty and a corresponding liability. Legal and ethical rights and obligations pursue different goals. Whereas the former aims to guide "the good deed," the latter's task is to maximize equal individual liberty. Hence, currently the law (not only in England) is unlikely to find surgeons liable for negligence when they take, or fail to take, mental enhancers.

Elizabeth Shaw addresses the controversial topic of brain interventions in rehabilition programmes for criminal offenders. At the outset, she presents potential aims of neurointerventions: increasing empathy, decreasing violent and sexual urges and even racial sentiments, strengthening willpower as well as enhancing cognitive abilities. If safe and effective means are developed in the future, a host of ethical and legal questions arises. Shaw addresses and ultimately dismisses the most prevalent ethical criticisms of "moral enhancement." In light of the law, interventions were particularly worrisome if they affected a person's status as a moral agent by undermining their free will. However, so Shaw argues, the interventions available today or in the near future do not globally weaken capacities for autonomy.

The most controversial – and as yet not thoroughly addressed – form of changing offenders are mandatory interventions. As part of their punishment, states might sentence offenders to undergo neurointerventions. Shaw enumerates several arguments for and against compulsive treatments. For one, in some jurisdictions sexual offenders are treated against their will, sometimes with profoundly invasive methods. Furthermore, neurointerventions might be a more humane treatment than incarceration (or castration, as the case may be), given the situations of prisons in some countries and the physical and psychological effects of long-term imprisonment. In addition, by drawing on John Rawls' thought experiment of the "original position" in which future citizens of the community to be (or rather their proxy) design social rules from an impartial perspective behind a "veil of ignorance," some theorists argue that consenting to such treatments might be in the best interest of offenders and can hence be normatively presumed. Others, more harshly, suggest that offenders have simply forfeited their rights against being subjected to neurointerventions. Finally, interests of third parties might be taken into account, such as potential victims, families, or the public (as imprisonment is exceedingly expensive).

However, strong arguments speak against mandatory interventions. Shaw highlights the risks of misuse if states were to acquire the power for invasive intrusions. She presents two arguments, based on a retributivist and a non-retributivist position on punishment that both provide strong reasons against mandatory interventions. In addition, the author reminds us that state's powers to punish are limited. The 8th Amendment of the US Constitution, for instance, prohibits "cruel and unusual punishment." In other jurisdictions, the inalienable right to human dignity might be affected by sufficiently invasive interventions because they may be viewed to objectify offenders. Furthermore, Shaw is critical of the often and all too easily drawn analogy between incarceration and neurointerventions. While the former

only restricts freedom of movement, neurointerventions encroach upon essential characteristics of the person which are often conceived of as being, for principled reasons, beyond the reach of governmental control.

In the face of these objections, the best way might be to offer neurointerventions on a voluntary basis. Besides some familiar questions of informed consent, the special context of prisons might, however, cast doubts on the "real" voluntariness of offenders' consent. Shaw engages with threats and inappropriate offers and draws distinctions between different aims of the intervention. Whereas experimental interventions in exchange for early release may take unfair advantage of the offender's situation, the state is entitled to offer interventions targeting the psychological sources of the behavior for which the offender was sentenced. For if the incarceration (as the only alternative to the brain intervention) is the result of the proper application of a legitimate legal order, it should not be considered analogous to (illegitimate) pressures from other people, but rather akin a necessitating situation originating from one's natural "bad luck." After presenting further arguments of the current debate, the author concludes that under certain conditions neurointerventions can and should be offered by the state as long as they remain respectful of the offender's status as a moral agent.

Cross-References

- ▶ A Duty to Remember, a Right to Forget? Memory Manipulations and the Law
- ▶ Cognitive Liberty or the International Human Right to Freedom of Thought
- ▶ Compulsory Interventions in Mentally Ill Persons at Risk of Becoming Violent
- ▶ Ethical Implications of Brain Stimulation
- ▶ Neuroimaging and Criminal Law
- ▶ Neuroimaging Neuroethics: Introduction
- ▶ Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges
- ▶ Responsibility Enhancement and the Law of Negligence
- ▶ Smart Drugs: Ethical Issues
- ▶ The Use of Brain Interventions in Offender Rehabilitation Programs: Should It Be Mandatory, Voluntary, or Prohibited?
- ▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

A Duty to Remember, a Right to Forget? Memory Manipulations and the Law

82

Christoph Bublitz and Martin Dresler

Contents

Abstract

Neuroscience might develop interventions that afford editing or erasing memories, changing their content or attenuating accompanying emotions. This section provides an introduction to the intriguing ethical and legal questions raised by

C. Bublitz (✉)
Faculty of Law, University of Hamburg, Hamburg, Germany
e-mail: christoph.bublitz@uni-hamburg.de

M. Dresler
Max Planck Institute of Psychiatry, Munich, Germany

Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands
e-mail: martin.dresler@donders.ru.nl

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_167,

such alterations, with a special focus on the report of the President's Council "Beyond Therapy" and the proposal of a right to freedom of memory advanced by Adam Kolber.

Introduction

Memory is a marvelous capacity. It allows representing the past and adapting future behavior in light of prior experiences. In a more colorful way, it is described as "mental time travel" (Tulving 1985) or "the presence of an absent thing stamped with the seal of the anterior (Ricoeur 2004, p. 17)." Memory is a multifaceted phenomenon, not a single entity but a bundle of functionally distinct yet interwoven capacities with different neurophysiologic foundations that enable the most basic forms of learning such as classic conditioning as well as reminiscing about the trajectory of one's life.

Current memory models differentiate several systems by content (LeDoux 2007): Declarative memory comprises semantic, autobiographic, and episodic contents that can easily be verbalized, whereas non-declarative memory stores motor or perceptual skills and conditioned stimulus–response patterns which can be behaviorally expressed but hardly articulated. Thus the verb "to remember" does not denote a single activity but comprises, e.g., the ability to express acquired procedural skills and factual knowledge as well as reexperiencing the past with that distinct phenomenal quality for which Elving Tulving coined the term "autonoetic consciousness" (Tulving 2002). In the following, we shall be primarily interested in the latter.

Further distinctions are drawn between temporal phases and processes. Memory content has to be acquired, stored, and recalled – accordingly, encoding, consolidation, and retrieval constitute the three main phases of memory. Directly upon encoding, memory traces are labile and prone to rapid decay. Most items in working memory (e.g., after perception) are never processed further; only some are transferred into long-term memory. During this process of consolidation, neural memory traces – also called "engrams" – are organized and stabilized. How memories are exactly stored remains a mystery and seems to differ for various types; the search for the engram is still on (Thompson 2005; Hübener and Bonhoeffer 2010). By all we know, memories are not stored in easily localizable brain structures but rely on networks of multiple brain areas and changes in the synaptic strength between neurons. The terms storage and consolidation suggest that contents of memory become permanently fixed traces in the brain, insensitive to modification, stored away until retrieval. However, such a static picture misses the dynamic and changeable nature of memories: During the retrieval process, sometimes induced through conscious effort, sometimes involuntarily, the engram seems to return into a temporarily fragile state and undergoes another process of consolidation (called reconsolidation) afterwards (Dudai 2012). Reconsolidation in humans is still subject to controversy and investigation and may have far-ranging implications (Schiller and Phelps 2011; Nader and Einarsson 2010). During the short period of reconsolidation, suspected to last

for a couple of hours, memories become open for alterations and seem to be "updated" in light of present knowledge. Reconsolidation might imply that the best recollection we have of a past event is our latest recollection of it (and not the original perception which may have been "overwritten" by the altered memory). As a consequence, our memories are alterable without our awareness. Memory is not, as the ancient metaphor of a wax tablet or modern variants of a computer storage device suggest, a replication or a copy of the initially encoded stimuli but a repeatedly reconstructive process and thereby susceptible to alterations and contaminations. A remarkable fact, because from our subjective experience, we are (too) confident that memory is more of a replay of recorded information than an active, potentially error-prone process.

Whoever speaks about remembering should not forget forgetting, an equally interesting and important process. Already its definition presents problems. As not every transient inability to recall (such as tip-of-the-tongue inhibitions) amounts to forgetting, how do we know, as Nietzsche pondered, whether we ever forget? Smells, sounds, or sights can trigger quite powerful memories which were irretrievable in the absence of such cues. Strictly speaking it can never be ruled out that some regularly inaccessible memory traces, excitable under the right conditions, still remain (Cf. Roediger III et al. 2010). Like remembering, forgetting is a multifaceted phenomenon for which no unified theory exists. It can have various causes: faulty initial consolidation, retroactive interference, trace decay over time, or simple retrieval inhibitions (Wixted 2004; Schacter 2002).

In any case, forgetting should not be conceived of as cognitive defect only. It is instrumentally useful for thinking – and possibly even remembering – because it prevents informational overload by filtering pieces worth remembering. Without forgetting we may suffer the fate of Funes the Memorious, a fictional character by Luis Borges, who cannot forget and is completely observed in details (Borges et al. 2007). Patients resembling Fuentes and suffering from their hyperthymestic memory exist (Parker et al. 2006). William James once remarked: "Selection is the very keel on which our mental ship is built. And in this case of memory its utility is obvious. If we remembered everything, we should be on most occasions be as ill off as if we remembered nothing" (James 2007, p. 680). Even apart from cases of complete memorizers, accurate recall is not always beneficial. One does not need to subscribe to controversial Freudian notions like repression or psychological defenses to acknowledge that too much of the past can burden the present.[1] Remembering all unpleasant or embarrassing moments is not conducive to well-being. Especially autobiographical memory seems to be selective. Attenuating or amplifying particular memories helps to maintain a coherent and positive self-image (Conway 2005). Reversely, goal-directed forgetting in the service of personal needs lets us see the past with misty eyes (Harris et al. 2010, p. 254). Remembering and forgetting are thus complementary processes and we should not overvalue one while depreciating the other.

[1]A contemporary review of repression theory; see Erdelyi (2006).

Interventions into Memory

We have to leave it here with this brief introduction to memory research and turn to what is relevant for normative purposes: interventions modulating memory. It goes without saying that our conscious powers over memory are limited. In recent years, several novel interventions have been studied that affect the fate of recently acquired memory traces and that might also afford altering more remote memories. Although most research is still in its infancy, the prospect that novel insights into the working of the memory systems and novel intervention techniques may allow to alter memories seems warranted in principle.[2] These are exciting times for memory researchers even though for safety and ethical reasons, many of the interventions studied in animals will never be carried out in humans.

Three main goals for interventions into memory are conceivable: enhancing memories through improving encoding and retrieval; diminishing memories through inhibiting recall, erasing traces, or attenuating emotional aspects; or changing contents of memories. Generally, interventions into memory are possible in all phases: before or during memory acquisition, consolidation, or recall – and, perhaps most interestingly, even after recall during reconsolidation. Moreover, case reports of patients with deficits in particular memory functions suggest that specific memory systems might be targeted individually.[3]

A growing body of research seeks to enhance memory by different means, mostly by improving information encoding or consolidation. A powerful natural way to enhance memory is emotional arousal which leads to increased release of stress hormones such as noradrenalin and cortisol that enhance encoding and consolidation but can interfere with retrieval (Joëls et al. 2011). Some common drugs have both memory enhancing and impairing effects, depending on the time of application: Benzodiazepines, for instance, impair memory anterogradely when taken before but enhance memory retrogradely when taken after encoding (presumably through reducing interference with later incoming information) (Beracochea 2006). The public debate on memory enhancement focuses mainly on pharmaceuticals, but most drugs currently in use in humans fail to exhibit overly impressive beneficial effects on memory (Repantis (2010a, b; Husain and Mehta 2011; Lynch et al. 2011). Several non-pharmacological memory enhancers have been demonstrated to be similarly or even more effective than drugs, e.g., nutritional supplements, physical exercise, sleep, mnemonic strategies, or brain stimulation.[4]

[2]There will be many obstacles, especially in targeting specific memories; see, e.g., Levy (2007), Ch. 5.

[3]The most famous patient is the recently deceased H.M., c.f. Corkin (2002).

[4]For a broad overview c.f. Dresler et al. 2012; for nutrition c.f. Smith et al. (2011), Nehlig (2010); for exercise c.f. Roig et al. (2013), Hötting and Röder (2013); for sleep c.f. Rasch and Born (2013), Genzel et al. (2014); for mnemonics c.f. Karpicke and Roediger (2008), Worthen and Hunt (2010); for brain stimulation c.f. Coffman et al. (2014), Suthana and Fried (2014).

Different strategies exist for diminishing memories. Forgetting is the often unintended but natural fate of most memories. There are some indications that intentional forgetting of selected memories might be possible, a phenomenon called “directed forgetting” (Geraerts and McNally 2008). For emotional memories like conditioned fear, unlearning procedures (i.e., fear extinction) have long been established and have attracted renewed interest recently. Notably, however, most fear extinction procedures do not erase memory traces but merely inhibit recall through a newly learned safety memory (Myers and Davis 2007; Vervliet et al. 2013).

A primary target for interventions are emotions that accompany memories, particularly painful and stressful emotions related to memories of traumatic events. In most persons who have experienced trauma, the emotional tone of the memory with time disconnects from the factual memory, a process in which sleep plays a crucial role (Walker and van der Helm 2009). Some persons, however, develop post-traumatic stress disorder (PTSD), a pathological condition characterized by anxiety, easily recurring and hard to suppress memories of traumatic events, flashbacks, emotional numbing, or hyperarousal, as well as behavioral changes such as evasion of particular people or places.[5] The powerful and persistent memories are caused by stress hormones released during the traumatic event which lead to an overconsolidation of memory traces (Pitman 1989; Glannon 2006). Most therapies aim to blunt the strong emotions associated with the factual content of the memory. However, insofar as the original memory trace is not replaced but only its recall inhibited, relapse is a common problem (Vervliet et al. 2013). Novel interventions might open new routes for prevention and treatment of PTSD. Substances could be administered to persons before they will be exposed to potentially traumatic situations (rescue workers) or immediately after they have experienced, e.g., sexual assault; or in therapeutic settings during reconsolidation of reactivated traumatic memories (President’s Council on Bioethics 2003; Donovan 2010; Poundja et al. 2012; Schiller et al. 2010). Consolidation and reconsolidation of emotional memories have been successfully targeted with the β-adrenergic blocker propranolol (Cahill et al. 1994; Kindt et al. 2009). In pilot studies, tampering with (re)consolidation processes has successfully reduced PTSD symptoms, but larger studies have yet to replicate these effects (Pitman et al. 2002; Vaiva et al. 2003; meta-analysis by Lonergan et al. 2013). Whether these interventions only diminish emotional aspects or affect factual recall is unclear at present.[6]

[5]For the biological side of PTSD, see Pitman et al. (2012).

[6]Some researchers speak of “erasing emotional memories” by which they mean erasure of the emotional aspect only, while ethicists often speak indiscriminately of “blunting” of painful memories. These conceptual ambiguities might lead to misunderstandings and depend on the conception of memory traces. Cf. Holmes et al. (2010). For present purposes, it is only important that emotional and factual contents could, at least to some degree, be targeted and modified independently from each other.

The strongest form of memory manipulation is complete memory erasure. The first point for intervention is blocking initial consolidation to prevent information in short-term memory from being stabilized and transferred into long-term memory. It is also possible to intervene at later stages: During the labile phase upon recall, reconsolidation could be impaired or prevented (Pitman 2011; Parsons and Ressler 2013). In animal studies, blockade of both consolidation and reconsolidation has been repeatedly demonstrated by inhibiting protein synthesis in memory-related brain regions (Dudai 1996; McGaugh 2000; Nader at al. 2000; Alberini et al. 2006; Shema et al. 2007; Sacktor 2008). In humans, reconsolidation of episodic memory has been disrupted, e.g., through emotionally aversive stimuli or electroconvulsive therapy applied immediately after retrieval (Strange et al. 2010, Kroes et al. 2014).[7]

Another form of intervention changes contents of memories. The process of selective consolidation of a fraction of our experience and the repeated reconsolidation of these memory traces suggests that our memories may be less veridical than we expect. This intriguing view is backed by a long line of false memory research, most notably by Elizabeth Loftus, in which the content of memories could be altered or new memories implanted through various and quite simple means such as telling false stories about the past or suggestive questioning (Loftus 2003; Brainerd and Reyna 2005). In animal studies, false memories were implanted through sophisticated optogenetic interventions at the molecular level (Ramirez et al. 2013). The extent to which our memories deviate from our original experience cannot be reliably estimated at present, both in terms of false parts within a (by and large) correct memory and the overall amount of false memories. In general, the possibility of false memories is not the same as proving that many are indeed false. Supposedly, humans are capable of remembering many things correctly. Otherwise, the preservation of our memory system through natural selection appears unlikely (Schacter 1995, p. 25). However, Loftus' studies demonstrate that distrust in memory is warranted and that we can never be sure whether our recollections are correct – regardless of how vivid and familiar they appear to us. These findings have legal implications, especially for evaluation of eyewitness testimony and interrogative procedures (Schacter and Loftus 2013; British Psychological Society 2008; Nadel and Sinnott-Armstrong 2012).

Ethics of Memory

The ethics of memory is a fascinating topic that has received little attention yet (Cf. Blustein 2008; Margalit 2002). How should we remember – what, whom, and

[7]The current state of research in humans is less consistent than in animals, partly due to the complexity of interactions between different memory systems during reconsolidation; for a review c.f. Schiller and Phelps (2011).

in which way? How often and intense should we, for instance, reminisce about late family members? To whom do we owe duties of remembrance, to ourselves or the person remembered? Should we try to forget particular persons or events or, by contrast, struggle against the natural decay of memories by fabricating cues and records? All these questions converge to the more general one: To which degree should the past influence the present? Answers have to take our limited powers over memory into account. But in a time when technologies may confer more control over remembering and forgetting and turn them into matters of choice, ethics of memory becomes an important issue.

The prospect of manipulating memory beyond our natural abilities has sparked the imagination of authors and artists from antiquity to modern days.[8] Although different interventions raise different questions and deserve detailed treatment on their own, some ethical worries apply to all of them (we loosely speak of "altering" memories). On the level of the individual, two ideas pull into opposite directions: For one, historical truthfulness speaks in favor of maintaining accurate recollections of the past. It finds support in prominent ethical ideas such as the ancient "know thyself." By contrast, well-being and the pursuit of happiness may favor altering memories – even forgetting unpleasant ones – over accurateness. In both cases, memory serves as an instrument for cherished but potentially conflicting ethical ideals.

A good example of the first position is formulated by the US President's Council of Bioethics in its report "Beyond Therapy." It presents a balanced overview of potential benefits and pitfalls of memory alterations and placed the topic on the scholarly agenda. It has nonetheless attracted much criticism for its bioconservative stance and its (rather suggestive) conclusions. The Council expresses deep concerns over any form of memory alteration, including interventions that numb traumatic memories of PTSD patients: "Our memories make us who we are. By 'rewriting' memories pharmacologically we might succeed in easing real suffering at the risk of falsifying our perception of the world and undermining our true identify" (p. 227).[9]

Remembering Fitly and Truly

As an ethical standard, the Council suggests "remembering fitly and truly" (p. 228), unfortunately without stating more clearly what "fitly and truly" means, what it implies or in which relation both criteria stand. "Fitly and truly" is presumably best understood as remembering appropriately and accurately and rules out falsifying,

[8]Cf. movies like Eternal Sunshine of the Spotless Mind.

[9]Cf. Parens (2010) who reads the Council as not being opposed to PTSD treatment.

forgetting, or dampening memories. Prima facie, this appears to be an agreeable position. But let us explore it a bit deeper. Any obligation to remember accurately faces the problem that human memory is selective and distortive. Emphasizing accurateness would oblige us to rehearse memories that tend to be forgotten and dissuade us from romanticizing or glorifying the past. Consequently, interventions that improve true recall and diminish distortive effects should be endorsed. This, however, stands in some contrast to the spirit of the report, critical of any technological alteration of natural human abilities.

Eric Parens usefully suggests understanding "remembering fitly" in the following way: The intensity of memories and accompanying emotions have to stand in some relation to the magnitude of the event that caused them (Parens 2010). This criterion brings us some way in clearly disproportional cases. However, assessing the significance of events as well as adequate (emotional) responses is itself a value judgment in need of further criteria. What, for instance, is the appropriate response to the death of a friend, the breakup of intimate relationships, or embarrassing experiences? Hard to tell. And by whose standards? (Cf. Henry et al. 2007, p. 17) This is particularly important in regard to reactions to traumatic memories such as deep prolonged sadness and loss of trust or interest in mundane things. These might be quite appropriate responses to preceding events such as having gone through war or being sexually assaulted. In fact, how could we simply revert back to normality and enjoy life's pleasures after living through such experiences? Even if affected persons were haunted by memories for years, their reaction is not evidently disproportionate to the magnitude of the event. The problem with traumatic memories is not so much that they are inappropriate to the past but that they may lead to a dysfunctional life in the future. "Fit" and "functional," though closely related, have different reference points: appropriate vis-à-vis the past or a flourishing life in the future.

This leads us to the third ethical ideal which a US American Council cannot but recognize: the pursuit of happiness. The Council contends that remembering fitly and truly is a precondition for living a flourishing life. But this is dubious – at least, but not only – with respect to traumatic memories. The Council's argument heavily draws on the idea that there is something like "real happiness," different from feelings of pleasure, which requires coherence between the world and subjective feelings. Indeed, we should be skeptical about superficial happiness, gained only for the price of separation from reality and may (mis)appropriate Theodor Adorno's famous saying here that there is no real living in a false life. But albeit distrust against superficial false consciousness is warranted, it is hard to get around the fact that certain memories are debilitating which is not a desirable state and that memory alterations could assist overcoming inner obstacles grounded in the past. The resulting flourishing life might not be so "unreal" or "false" that it is not worth having. While the Council claims that happiness and remembering fitly and truly are parts of the same coin, we suspect they are different ideals that can come in conflict with each other. Out of this tension, the real challenge arises: Are we obliged to remember fitly and truly *even though* it impairs a flourishing life?

Identity

In the background of the argument stands the worry that memory alterations threaten identity. Indeed our autobiographical memories make us who we are. If taken away from us, we would be stripped from something fundamentally ours, deprived of some kind of access to ourselves – our autobiography. Nonetheless, the importance of every single memory should not be overstated. We all forget a myriad of things and events on natural ways without loosening our identity. Why should this be qualitatively different if we forget at will? Furthermore, a well-considered decision to erase memories might itself be an expression of one's personality.

Worries over identity are a common topic in bioethical debates. It has proven useful to differentiate between several meanings. The strongest form, diachronic identity, concerns the continual existence of a person, i.e., the conditions under which a person at one point in time can be reidentified as the very same person at a later point. In the wake of John Locke, some argue that autobiographical memory, experiential knowledge of one's past, is a necessary condition for personal persistence over time.[10] This implies that loss of autobiographical memory leads to a fissure in persistence: The new person is, in a strong sense, a different person, while the old one has vanished.[11] Accordingly, Alzheimer's patients suffering from irreversible memory loss would be numerically distinct from the persons that inhabited their bodies before onset of the disease. While the tragedy of dementia-related erosion of personality is indeed a striking example of memory's importance, it also points to a weakness of the memory criterion. Theory aside, we usually identify the old lady who cannot remember her past as our grandmother who raised us as kids and therefore feel (and are) obliged to care for her.

As any theory of diachronic identity has to allow for the ordinary degree of forgetting, interventions such as numbing or erasing particular memories will regularly not call it into question (although the Council suggests otherwise). Yet, a weaker sense of identity, closer to the colloquial use, might be affected: one's personality. Everyone has a past, and this past cannot be undone. If memories of our past make us who we are, changing our memories may change us. The question is whether altering our personality in this way is wrong. Essentialist thinkers urge that one should preserve rather than "betray" who one is. By contrast, existentialist-minded positions hold that we should be authors of our life, actively shaping our future in light of attractive self-images, even if that implies radical departure from former personality traits (Cf. Bublitz and Merkel 2009; Erler 2011). The wrongness of "betraying" who one *was* in the quest of creating a more desirable future version of oneself can hardly be found in claims of the former personality against its successors. Instead, it needs to be grounded in present interests of the person. In their pursuit, anyone attracted to

[10]In Locke (1979), he wrote of the "sameness of consciousness," traditionally understood as the memory criterion, but see, e.g., Strawson (2011).

[11]For more elaborated treatments on the relation between identity and memory, see Parfit (1984), DeGrazia (2005), Schechtman (2005), Galert and Hartmann (2007).

existentialist approaches may seek to alter the way she reacts to her past, including her memories, and concede historical untruthfulness for the sake of personal development. Abandoning parts of one's past is not tantamount to abandoning any personality, but altering it. Objections are thus convincing only if self-development is intrinsically wrong or if freeing oneself from one's past is counterproductive to that end. Given that many people seek to change the courses of their life and their way of being, often for the better, a wholesale rejection of self-creation appears implausible.

Self-Growth

Perhaps, shaping one's future through cutting links to one's past is doomed to failure. We take this to be primarily a psychological hypothesis. Without doubt, coming to terms with oneself and one's past is among life's major challenges. But might forgetting or editing memories not be part of it? Especially theories of "narrative identity" emphasize the constructive processes in personality formation. Persons compose stories about themselves, woven from various sources such as their recollections, beliefs, self-image, perception by others, and expectations and aspirations for the future (Cf. Galert and Hartmann 2007). Interestingly, studies suggest that persons tend to evaluate their past not truthfully but in ways conducive to well-being, by embellishing or depreciating it. Putting one's former self in a negative light might – just as downward social comparison – make one feel better about the present (Wilson and Ross 2001). Forgetting, selective remembering, and waning emotional reactions are thus to some degree ordinary processes that do not necessarily impair construing meaningful and functional narratives about oneself – they even seem to be regular features (Bell 2008). Then, truthfulness to the past might not be a central condition for self-development.[12] While this does not imply that editing memories is ethically advisable, it shows that it is not an insurmountable obstacle for self-development.

Nonetheless, those who cannot draw upon their history might indeed be bound to repeat it. This old wisdom seems to have a neuroscientific analogy: Findings suggest that persons with memory deficits also have difficulties in prospective planning because the same brain systems are involved in both tasks. Scientists speculate that the (evolutionary) aim of remembering is not accurate reproduction of the past but rather simulation of the future.[13] Be this as it may, engaging with one's past can be extremely beneficial for personality development. It helps to understand where one comes from and enables learning from (unpleasant) experience and developing strategies to deal with stressful events and ordinary nuisances of life. Accordingly, when we lose our history, we lose opportunities for self-growth. This may even be true for

[12]Some memory disorders impair forming a sense of self, e.g., Klein et al. (2004).

[13]The "constructive nature of episodic memory is attributable, at least in part, to the role of the episodic system in allowing us to mentally simulate our personal futures," Schacter and Addis (2007), p. 779, Schacter et al. (2007).

traumatic events which can acquire meaning in retrospect. Some traumatized persons experience "posttraumatic growth" and steer their life in novel directions (Calhoun and Tedeschi 2006). Blunting or erasing memories may block such developments. Yet, even though people give meaning to horrible events, it remains doubtful whether there is meaning to, e.g., being the victim of serious crimes. Presumably, giving meaning to tragedy is more a strategy for coping with emotional turmoil. Painful events should not be glorified because of potential positive side effects. Memory alterations could provide alternative ways of overcoming the shadows of the past and freeing up the inner resources that arduously working through trauma exhausts. But while therapeutic forgetting may have much to recommend, concerns that memory alterations could be used to overcome minor troubles and impair personal growth seem warranted, especially in light of notorious human traits such as impatience and discounting long-term drawbacks in favor of short-term benefits.

Finally, it should be reminded that even full memory erasure cannot ensure escaping one's past as long as others retain their recollections.[14] The fact that others know more about a person than she herself does can lead to bizarre and uncomfortable situations which might be worse than retaining the original memory. And we can learn from patients suffering from involuntary memory loss that many persons would feel urged to find out as much about their history as possible. Then, memory erasure would be self-contradictory.

After all, the question about the extent to which the past should influence the future touches upon an amalgam of potentially conflicting ethical ideals, none of which can claim strict priority. Remembering fitly and truly is neither conceptually nor empirically a necessary condition for a flourishing life. Corresponding normative demands may speak in favor of manipulating natural abilities which do not live up to those ethical ideals. Nonetheless, memory alterations pose various dangers, many uses appear imprudent and should be discouraged. In spite of this, there are potential benefits. In the end, much depends on the kind of memory altered, precise effects of interventions, as well as psychological and social consequences. These aspects cannot be determined a priori but have to be cautiously explored empirically.

Memory & the Law

The perspectives of law and ethics differ in important ways. The law does not provide answers about what is morally advisable to do. Unlike moral advice, legal provisions are binding for all. Neutral and pluralist democracies should not seek to impose on citizens contested views of a good life but enable the peaceful coexistence of diverging life plans. Specific legal provisions have to conform to higher level principles that give form to the structure of legal norms, primarily those deriving from constitutional and human rights law. Deeply entrenched in liberal

[14] In-depth discussions of further ethical aspects are Levy (2007) and Liao and Sandberg (2008).

rights is the idea that persons enjoy wide ranging autonomy in self-regarding matters, which entails the liberty to make imprudent and immoral decisions. The realm of legal freedoms is thus not coextensive with the morally good. Because of this, legal paternalism is highly contested, at least in theory (see, e.g., Feinberg 1986). In practice, most legal systems allow hard paternalism in order to prevent severe self-harm. Furthermore, legal rights are understood here as prima facie entitlements that have to be balanced against countervailing rights of others or legitimate public interests before final judgments can be made. Finally, since legal systems and cultures differ widely, the following has to remain on an abstract level and might not apply to every jurisdiction to the same degree.

In most countries, the use of memory-altering substances or tools is regulated by different legal provisions. Pharmaceuticals such as propranolol or antidementia drugs are scheduled substances which require prescription. Tools like brain stimulation devices fall under different regulations, while mnemonic training, physical exercise, meditation, or sleep are not regulated at all. Many provisions that restrict access to specific means are not enacted in virtue of their effects on memory but for other reasons. Current regulations based on means may thus appear incoherent sub specie memory and might be reexamined in view of present knowledge.

In general, only few legal provisions directly pertain to mental states. At least in theory, strong human rights such as freedom of conscience and thought place some parts of the mind outside the reach of governmental control.[15] Memory, however, is among the exceptions. States have always had legitimate interests in acquiring knowledge about the past. Every citizen can be summoned to testify as a witness and this may entail the duty to remember correctly.

Two rights are noteworthy: Mental health, a human right, e.g., under the European Convention of Human Rights,[16] affords citizens claims against states to refrain from actions that inflict mental harm and to protect them against such actions by private parties. Thereby it provides protection against infliction of memory disorders that amount to mental health problems. In addition, curbing access to effective (and relatively safe) treatments of mental disorders would interfere with the right to mental health. Consequently, it appears unlikely that courts would consider the promotion of ethical ideas such as truthfulness to one's past as sufficiently grave interests to justify, e.g., prolonged suffering from PTSD.

Moreover, some aspects of memory concern issues of identity and personhood and may therefore relate to human dignity.[17] For instance, retrograde amnesia or personality dissolving effects of advanced Alzheimer's can undermine dignity (make a dignified life almost impossible). Respect for dignity not only implies the obvious – a strict prohibition to contribute to severe memory loss – but might

[15]Art. 9 ECHR, Art. 18 Universal Declaration of Human Rights.

[16]Art. 8 ECHR, e.g., Bensaid v. UK, App.No: 44599/98, 6.5.2001.

[17]Although not always codified and subject to controversy, human dignity is often understood as the overarching principle of human rights law.

also oblige states to prevent the onset of these diseases, e.g., by providing access to antidementia treatment. However, the extent of positive obligations in general and to health care in particular varies greatly from country to country.

Freedom of Memory

Interventions into memory not directly related to dignity or health might not be adequately captured by current law. Defining novel, memory-specific regulations could soon become a challenge for lawmakers and legal theorists. As a starting point, Adam Kolber has proposed to acknowledge a novel right, "freedom of memory," which he describes as a "yet poorly defined bundle of rights to control what happens to our memories" (Kolber 2006, p. 1622; Kolber 2008). We concur with Kolber's proposal. Freedom of memory follows straightforwardly from the general presumption of liberty and is part of the special protection of the person, arguably the foundational concern of human rights law. In the following, we shall give some contours to the idea and sketch issues future scholarship has to address.

Scope of the Right

Right to Remember

Our strongest interest in memory is remembering our past, to preserve biographical events against sinking into oblivion. In virtue of its significance, it deserves heightened legal protection. On a basal level freedom of memory thus entails the right to remember.[18] It guarantees that persons are entitled to use their powers of memory at will. This right corresponds with a duty of others to refrain from interfering with memory. Interventions that impair memory such as electroconvulsive therapy or pharmaceuticals interfere with this right and require strong justification if administered without consent. We suggest, as a rule of thumb, that the right to remember is stronger – and more likely outweighs countervailing interests – the higher the relevance of particular memories or capacities for construing a meaningful and truthful life narrative.

In addition, the right has to protect accurate memories against distortive influences. Courts were confronted with this issue in repressed memories cases in the 1990s. According to some psychological theories, repressed traumatic childhood experiences can be recovered through explorative psychological procedures. Upon undergoing such treatments, some patients remembered that they were sexually abused as children and brought claims against ostensible perpetrators. These memories often turned out to be false. They were not recovered but (negligently) implanted by psychotherapists through suggestive techniques such as hypnosis, guided imagery, drugs, and positive feedback for reports of memories with abusive

[18]A right to remember does of course not entail a duty to remember.

content (Cf. Loftus and Ketcham 1996; Brainerd and Reyna 2005, Ch. 7). At times patients recalled bizarre events such as satanic ritual abuse and group rape. More astonishingly, even some alleged perpetrators confessed horrible crimes which they never committed, only because suggestive police interrogations implanted false memories.[19] For lack of reliability and scientific consensus, courts grew reluctant to admit evidence of repressed memories. Today, chances of conviction based on testimonies from recovered memories without further corroborative evidence are low.

These cases raised the question whether patients or the wrongly accused have causes of action against therapists. The patient–psychotherapist relation is subject to contract law; therapists have to provide treatment *de lege artis*, according to standards of medical practice, which were arguably not observed in these cases. At least, methods that potentially alter memory should, just as physical interventions into bodies, require informed consent. Patients have to be informed about the scientific status of the repressed memory paradigm and the risk of false memories. Implanting false memories may thus give rise to malpractice suits.

Moreover, the work of Loftus demonstrates that false memories can be implanted outside of psychotherapeutic contexts. To capture such cases, the law would have to establish noncontractual duties of care toward other persons' memories. The problematic point is that false information can suffice to distort memories, but as misinformation is virtually everywhere, it can by itself hardly warrant tort claims (Morgan et al. 2013). The law has to tailor more narrow duties for persons with special responsibilities such as interrogating police officers. Implanting memories not as innocuous as the one's used in research, e.g., having been lost in a mall as a kid, should suffice to ground tort claims (e.g., infliction of mental distress, a tort accepted in some but not all jurisdictions).[20]

Protection Against Erasure

Furthermore, the law should provide protection against unwanted memory erasure even when it does *not* lead to mental distress. Walter Glannon discusses a real case in point, here in a slightly modified form[21]: For removal of suspicious tissue, a patient has consented to local anesthesia and, if necessary, for full sedation. During surgery, the locally anesthetized patient overhears the pathologists diagnosing "bad cancer" and starts to panic. The doctor injects her propofol, sometimes called "milk of amnesia" because it induces short-term anterograde amnesia through blocking initial memory consolidation. After waking up, the doctor tells the patient the surgery has gone smoothly and reveals her devastating diagnosis in the following days.

---

[19]In a famous case, the accused confessed the murder of 25 infants after authorities pressured him to remember the events; cf. Levy (2007).

[20]A further question is whether third parties – the accused – can bring claims against therapist. In the landmark case Ramona v. Isabella, the court granted a wrongly accused father remedies; see Mullins (1996).

[21]Glannon (2010), 240f. The case is taken from TIME, Oct. 15th 2007.

This is a case of genuine memory erasure, but instead of causing distress the intervention temporarily relieves it. Similarly, perpetrators could erase memories of victims so that they are unable to identify them which could have the positive effect of preventing PTSD. Without causing mental distress, these interventions might not fall within the scope of currently accepted torts. In that case, tort law should be expanded to outlaw unwanted memory erasure regardless of its negative or positive consequences. Because of their importance, lawmakers may even consider to render the unwatend erasure of memories into a criminal offense (Bublitz and Merkel 2014). Of course, memory erasure might be justifiable in exceptional cases (arguably in the one discussed by Glannon).

Right to Enhance Memory

In the 1950s, Wilder Penfield pioneered experimental stimulation of the brain. Electrical stimulation of the temporal lobes evoked vivid memories of events long forgotten. At least, this is what patients reported. As those memories were never verified, they might have been mere fantasies (Schacter 1995, p. 12). However, more precise stimulation and better insights into memory traces could afford new ways to explore buried memories, and, so we suppose for the sake of argument, sufficiently veridical ones. Should persons have a right to undergo such procedures? Other means that potentially enhance memory such as the nontherapeutic use of antidementia drugs or mnemonic training raise structurally similar questions. Given widespread complaints about the fallibility of memory and the natural curiosity about one's past, a high demand for effective memory enhancements can be expected. As part of the struggle against episodes of one's life fading away and becoming irrecoverable, the use of enhancements falls within the scope of the right to remember. It is neither conceptually nor normatively confined to our limited natural powers of recall. It guarantees remembering as one pleases which implies a permission to employ memory-aiding tools. Restricting access to such tools interferes with the right.

Limits

These interferences could be justified. As a form of cognitive enhancement, memory-boosting tools face many of the ethical objections reviewed elsewhere in this volume. Apart from issues of safety, efficacy, and undesirable consequences of overly perfect memories, ethical and social concerns may justify limiting freedom of memory. However, in light of the importance of memory and the fact that our natural powers are limited – at times, tools may be the only way to acquire access to one's past – countervailing interests would have to meet a high threshold. Social interests strong enough to outweigh, e.g., the interest to remember significant events of one's life are hard to imagine.

Furthermore, restrictions of modern versions of the ancient *ars memoriae* such as mnemonic training seem hardly justifiable, although they may e.g., give a competitive edge in the job market. Access to pharmaceuticals can be regulated in virtue of side effects, but unless considerable negative effects on individuals or society at large are expected, doctors should be free to prescribe them. The liberty

to use memory enhancements becomes even stronger insofar as they have preventive effects against memory decay in the future.

In extraordinary cases, freedom of memory might find its limits in legitimate interests of others that some events be forgotten (Kolber 2006). Think about an offender's vivid memories of how he humiliated and abused a victim. Victims may have the understandable wish that these memories shall not persist. Yet, these memories may also be important for offenders (e.g., for coming to terms with their deeds). If means to erase particular memories become available, the law needs to strike balances between these interests. Apart from such extreme cases, the right to remember regularly prevails over social interests to forget.

Right to Forget

As memory is the interplay between recall and forgetting, the freedom of memory also entails the right to not remember and even to forget. Again, good ethical reasons may speak against intentional forgetting, but legal freedoms do not coincide with the realm of the morally advisable. Only severe self-harm can be prohibited on paternalistic grounds.

Not Being Reminded of Something

A right to forget entails that others do not have claims against the rightholder to remember. Therefore, the romantic promise of lovers to never forget a precious moment (and each other) cannot be understood as a binding and enforceable legal contract. Moreover, in our quest to forget, we try, to suppress thinking about particular persons or events. Often in vain, as we are inadvertently reminded of them by external audiovisual or olfactory cues. The right to forget cannot protect against being reminded as this would restrict other persons' freedoms, e.g., the right to speak about issues others are not pleased being reminded of. Only particular means of evoking memories in others can run afoul of freedom of memory such as unwanted electrical stimulation of the brain. Unlike speaking, no one is entitled to directly interfere with another person's brain. Even if such interventions were free of side effects, the elicitation of unwanted memories by itself violates freedom of memory. Between speech and brain stimulation lies a grey area of ways to stir up another person's memories, e.g., through placing cues in public places. But if elicitation of memories were sufficient ground for banning cues from the public, all kinds of stimuli would have to be removed. The public sphere, however, attended to by different persons with different sensitivities, must remain a place for free expression even if some take offense. A different conclusion might be warranted in special cases, such as a graffiti artist who intentionally places symbols that trigger traumatic memories in the social environment of abuse victims (such cases do exist). It is not the illegitimate appropriation of the public for personal purposes but the intentional elicitation of memories that can ground claims in those cases. In general, PTSD and stressful memories are a growing field of litigation, and the law seems more willing to accept such claims, providing some protection for peace of mind, although scope and limits remain to be worked out in detail (Shen 2012).

Intentional Forgetting

While freedom of memory cannot protect against exposure to any memory cue, it encompasses the right to forget through straightforward measures. Umberto Eco once pointed to an asymmetry between the *ars memoriae* and *ars oblivionalis* (Eco 1998). While we can establish associations and cues to remember, there is no analog method to forget. Although some indirect strategies to suppress unwanted memories exist (Anderson and Green 2001), forgetting is, by and large, not under our conscious control. The advent of consolidation-blocking substances could change this, and their use would fall under freedom of memory. But, just like any other freedom, it can be limited by countervailing interests.

Paternalistic Limits: Erasure of Autobiographical Memory

Let us briefly turn to one example of justifiable paternalism: complete erasure of autobiographical memory. It may sound like mockery to the ears of those who suffer from severe memory loss, but some persons may seek to completely eradicate their autobiographical memory. At present, there are no means to do so, but tragic examples of brain lesions prove that it is possible in principle. Suppose an intervention erases memory without impeding the capacity to store new information. In the great majority of cases, availing oneself of such means would be an imprudent decision that causes more harm than relief and states should step in. However, cases are conceivable in which even this radical memory modification appears as an understandable wish, all things considered, e.g., if tragic circumstances have come over the person which were not of her own making and severely obstruct a flourishing life in the future. To some, autobiographical memory erasure disrupts diachronic identity and it might be conceived as a minor form of suicide.[22] Countries that tolerate or assist suicide might, in order to prevent suicide, discuss regulations and procedures for such interventions, bizarre as they may appear at first glance.

Limits: Collective Memory

The tempting short-term relief provided by memory modifications might come at the price of negative long-term effects on personal development for which paternalistic restrictions seem warranted. We cannot delve deeper into this contentious issue here and shall turn to limits which derive from interests of the common good. A recurring theme in the debate concerns the effects of memory alterations on collective memory. The Council argues that "our own memory is not merely our own; it is part of the fabric of society in which we live" and pictures the social consequences of memory dampening by asking: "What kind of people would we be if we did not want to remember the Holocaust, if we sought to make the anguish simply go away?" (p. 231) The Council concedes that "we cannot and should not force those who live through great trauma to endure its painful memory *for the benefit of the rest of us*." Yet, rather

[22]The law would probably not accept changes in numerical identity and still consider the person as the same as she was before erasure of autobiographical memory with respect to, e.g., financial obligations. Whether persons can still be punished for crimes committed before is a more intricate question, see, e.g., Dufner (2013).

than concluding that victims should be free to forget painful events, the Council suggests that society should compassionately suffer with victims instead.

The Council worries that atrocities and outrageous crimes could simply be forgotten and that victims would no longer seek to redress injustice. In that case, we would indeed be morally bankrupt. However, if we listen to the stories of (traumatized) survivors, the assumption that they could have chosen to forget if they only had effective means at their disposal seems misplaced. On the contrary, many seem to have derived the strength to survive from what they conceived as their moral duty: to tell the world what has happened and to seek justice, in the name and in memoriam of those brutally murdered. The engraving on the Holocaust Museum, “for the dead and the living, we must bear witness,” seems to express a deep existential commitment rather than an externally imposed duty. In the Oath of Buchenwald, survivors declared “it was one thought that kept us alive: the time for revenge will come” and pledged to “take up the fight until the last culprit stands before the judges of the people.”[23] For many, forgetting has never been an option.[24] However, we shall refrain from speculating about personal motivations of survivors because atrocities might simply not be the right angle for framing general discussions of memory alterations. How to come to terms with singular historical catastrophes far exceeds present parameters, and such events may not lend themselves to generalizations. We suppose that one can unconditionally endorse the imperative to never forget the Holocaust but still argue about manipulating memory in minor cases. Those we shall have in mind in the following.

The Council contends that memories are not “merely our own”, not at one’s free disposal because they are part of collective memory, and, further, that forgetting or numbing could impair or render impossible fulfilling collective duties of remembrance. However, the contested, perhaps only metaphorically useful concept of collective memory seems to be misleading.[25] Individual memories are the sources from which communities weave the stories about their past. The construction of these social narratives parallels individual memories in interesting ways: It pursues goals, primarily strengthening collective identities, not necessarily historical truthfulness. Parts of history are embellished, others are left out; stories are retold and rearranged through cultural and commemorative events in the service of present interests. Accordingly, historians speak of memory distortions on the collective level, too (Schudson 1995). Normatively, however, the fact that individual memories are sources of collective remembrance does not make them “memories of the community,” nor do communities derive claims to access or preservation simply from the fact that they can utilize them. Arguments in this vein presuppose that

[23]Oath of Buchenwald, April 19, 1945.

[24]How to reconcile the duty toward memory with the need to forget is a central theme for Holocaust survivors; cf. the writings of Elie Wiesel, e.g., his Nobel lecture, Dec. 11, 1986 (www.nobelprize.org).

[25]See Ricoeur’s (2004), p. 120 discussion of Halbwachs’ ideas who coined the term collective memory.

individuals are obliged to contribute to forming collective identities and facilitating collective goals. This might be a moral duty, but not a legal one.

Regarding duties of remembrance, an ambiguity of the concept "remember" might be at play. Collectively remembering historical events is not the same as individuals reexperiencing the past. The former means, e.g., drawing lessons from history, paying tribute to involved persons, and commemorating the dead. It does not mean, by contrast, that the collective or any individual entertains an autonoetic, first-person memory. The collective duty is independent from, and can be fulfilled without, autobiographical recollections.[26] Because of this, collective memories of historical events can be kept alive even when all contemporary witnesses have passed away. Erasure of individuals' memories does not automatically affect collective memory. Surely, oral accounts of moral witnesses can support collective remembrance, and they may have a moral duty to testify. But if witnesses cannot, as the Council concedes, be compelled to painfully remember for the benefit of society, their testimony should be encouraged but ultimately left to their discretion.

Social Interests: Witness Testimony

In this regard, the law is stricter. Irrespective of painfulness, it imposes duties to remember on witnesses for the benefit of the common good, more precisely, for purposes of law enforcement.[27] In many jurisdictions, the failure of a summoned witness to testify accurately constitutes punishable offenses. The law has to rely on witnesses. For lack of other obtainable evidence, testimony plays an important part in various legal proceedings. Memory alterations could impair the accuracy of testimony, fact-finding, and the administration of justice. For reasons of a fair trial, testimonies have to be given before a judge or a jury and be called into question by all parties. This makes oral proceedings necessary. As they unavoidably commence some time after the incident in question, the law has to address the duties of witnesses and the legal status of their memories during the meantime.

While the legitimacy of the duty to testify is hardly disputed in principle – everyone has to contribute his share to maintaining a (just) legal order – its precise contours are less clear. Before court, witnesses have to testify to the best of their knowledge. However, if memories were erased beforehand, witnesses cannot but truthfully refer to their memory blanks. We are all too aware of the notorious memory gaps that strikingly often befall people as soon as they enter courtrooms (not all of them bona fide). Since memory blanks can be beneficial for witnesses, they might be tempted to bring them about, and other interested parties might offer incentives to do so. Already today, witnesses might be able to render their testimony useless for legal purpose by exposing memories to ordinary risks such as alcohol or repeated recall under distortive influences (or seek the help of false memory implanters). At least intentional memory

[26]The same can be said about worries to "make the anguish go away". Is it the anguish of the victim that shall persist or not rather the anguish of noninvolved persons, as an empathic response to the suffering.

[27]In special cases, procedural rules recognize painfulness of memories, e.g., in sex-related cases.

distortions contradict the duty to testify truthfully. As self-induced witness amnesia would seriously obstruct the administration of justice and cannot be tolerated by the law, memory alterations that undermine accurate testimonies should be prohibited. Insofar as current provisions do not capture self-induced memory loss, offenses such as destruction of or tampering with evidence should be extended to encompass engrams in the brain.[28] However, tampering with engrams is in many ways dissimilar to tampering with physical objects. Whereas the latter can be confiscated and stored, persons can, in view of the dynamic nature of memory, hardly be obliged to preserve unmodified memories (or the initial perception). Some modifications are inevitable as long as memories are recalled and reconsolidated. Thus the law should reconsider the scope of witnesses' duties of care towards their fragile memories in light of current research.

Memory blunting might be the most pressing issue. On the empirical side, the impact of emotional blunting on factual recall has to be investigated. If emotional responses can be numbed while facts are largely preserved, the main legal interest in memory – gaining knowledge about the past – is not seriously threatened.[29] If, by contrast, blunting deteriorates factual recall, delicate balances have to be struck between societal interests in preserving accurate memories and witnesses' right to mental health. Roughly, we suggest that persons not responsible for the event such as victims or bystander witnesses cannot be expected by law to suffer intense and debilitating trauma. The short time window for interventions during consolidation (a couple of hours) raises a host of practical problems that need to be addressed (Cf. Kolber 2006, p. 1587).[30] Eventually, after witnesses have testified, the law does not stipulate further claims over their memories.

Effects on Responsibility

Another objection pertains to the effects of memory alterations on moral responsibility. The Council voices the concern that the idea of moral responsibility could unravel because victims would not seek justice any longer: Instead of forgiving, they would simply forget.[31] From a legal perspective, this worry seems farfetched and reverses the interests in question. Among the major aims for holding persons responsible is redressing harms inflicted on victims. The victim's main claim is restoration of the status quo ante, i.e., to put him in a position as if the harming

[28]See Kolber (2006), p. 1589 for a US-specific argument to this end.

[29]On occasion, the law might have further interests in unmodified emotions (e.g., in assessing damages), Kolber (2006), p. 1592.

[30]Perhaps immediate, taped interrogations to record at least a first unaltered testimony should be developed. But even then some parties have a disadvantage as they cannot cross-examine the witness. Also, in the immediate aftermath it is impossible to predict who will develop trauma, so the interests to be balanced are unclear in the moment in which actions have to be taken. Furthermore, it is often not evident whether memories will be of legal relevance; how long should people wait?

[31]Ricoeur (2004) argues that forgiving can be facilitated through forgetting.

action had never occurred (restitutio in integrum). Because history cannot be undone, this claim often remains unfulfillable, and secondary (financial) remedies are awarded instead. Easing negative mental consequences for victims would be the closest approximation to full restoration. Mitigating psychological harm, if necessary through memory alterations, is therefore a demand of justice, not its abandonment.[32]

The sinister consequences memory numbing may have on responsibility are often illustrated by two powerful images: One is the soldier with subdued emotions, blind and unresponsive to the horrors in which he participates, an epitome of a soulless human killing machine. The other is borrowed from Shakespeare's Macbeth:

> Cure her of that: Canst thou not minister to a mind diseas'd; Pluck from the memory a rooted sorrow; Raze out the written troubles of the brain; And with some sweet oblivious antidote Cleanse the stuff'd bosom of that perilous stuff / Which weighs upon the heart?[33]

Memory & Conscience

The idea of people washing away their feelings of guilt along with their recollections or of murderers with a clean conscience is highly troublesome. Accordingly it is widely held that perpetrators of crimes should not, as Macbeth hoped for his wife, numb their pangs of conscience through memory-blunting drugs (Parens 2010). On a closer look, memory plays only an instrumental role in these cases. What is really at stake is whether persons are sometimes obliged to experience feelings of guilt. Moral intuitions suggest that a guilty mind should be a conscience-struck mind. But is this also a legal duty? This intriguing question falls within the ambit of another, firmly established human right: freedom of conscience, enshrined in every human right treaty and among the core guarantees of the human rights system. Judicature and scholarship consider the inner side, the so-called forum internum where conscientious beliefs and emotions are formed, as protected unconditionally, i.e., intrusions are strictly prohibited. The central idea is that the conscience of the individual must, in principle, remain outside of the reach of governments. Although the precise contours of the right and its limits have yet to be defined,[34] freedom of conscience seems to imply that states do not have any legal claim over the conscience of the individual and that regulations of substances cannot be grounded in undesirable effects on conscience. The social interest in sustaining pangs of conscience then appears insufficient to curb its freedom.

[32]Furthermore victims have moral claims against perpetrators, e.g., to explain their reasons for actions. Memory erasure could thwart these obligations, which are often not enforceable by law because defendants enjoy privileges against self-incrimination and can remain silent.

[33]Shakespeare, Macbeth, Act V, Scene 3.

[34]The best analysis is Hammer (2001).

However, one should recall that the notion behind the absolute nature of the right lies in the protection of the individual against persecution for deeply held moral convictions and pressure to renounce one's faith. The origins of the right date back to confessional wars in medieval times. Neither in its historical genesis nor in legal scholarship the possibility of intentionally silencing one's "inner guiding voice" of conscience has been thoroughly considered. It raises the deeper question whether liberal states can require citizens to possess a minimum set of socially desirable psychological traits. Can a legal duty to be a moral agent, potentially riddled by a guilty conscience, be imposed on citizens; might it be legally enforced through denial of access to conscience-numbing drugs or even involuntary enhancement of moral dispositions? Perhaps. These issues run much deeper than freedom of memory. Current interpretations of liberty of conscience prima facie speak against a legal duty to be struck by one's conscience.

Memories of Perpetrators

As a consequence, soldiers are not legally obliged to feel remorse and cannot, in virtue of this fact, be stopped from altering their conscience, but military codes of conduct could proscribe use of memory-impairing substances. In practice, however, the military likely encourages numbing if it lives up to the promise of reducing the high rates of personnel suffering from PTSD. We share worries of soldiers killing with a clean conscience and – although we acknowledge the general weakness of arguments based on human nature – cannot but agree that some valuable essence of what it means to be human might be lost in those cases. However, we also sense some hypocrisy in the contention that soldiers ought to bear painful trauma for what others have commanded them to do.[35]

The most convincing case for an exception to the absolute protection of conscience can be made with regard to culpable offenders. In an insightful analysis, Carter Snead shows that the various aims of criminal punishment in one way or another presuppose that offenders and society remember fitly and truly. Memory modifications may thwart these aims (Snead 2011). What follows from this? In general, offenders do not have to contribute to the aims of the criminal justice system. This point is particularly important with respect to rehabilitation, the reduction of criminogenic psychological factors to reintegrate offenders upon release, which is among the main penological aims. In many countries, offenders are entitled to access to rehabilitative support but cannot be coerced into participation. States may provide incentives and rewards (e.g., reduced sentence, therapy as a condition for parole), but offenders can ultimately refuse rehabilitative attempts to remold their minds and serve their full sentence instead (and may

[35]Kolber (2006), p. 1621 draws the parallel between physical and emotional wounds. If treating the former raises no worries, why the latter? We suspect because of an (implicit) intuition that soldiers deserve mental torments, which is unconvincing as a general principle.

remain in preventive custody subsequently).[36] Whether coerced rehabilitation is viewed favorably depends on one's stance on the legitimatory grounds for punishment. The venerable issue may attract renewed attention if novel means for rehabilitation prove effective. Perhaps states might deny access to substances that subdue emotions conducive to rehabilitation such as guilt, remorse, or shame as part of punishment. Nonetheless, we hasten to remind that offenders are entitled to basic rights such as mental health, including treatment of painful mental consequences which originate in their wrongful deeds. Thus borders between self-incurred (and morally deserved) mental turmoil and pathological trauma might have to be drawn. While working through and coming to terms with one's deeds should be the default position, in cases of severe PTSD symptoms, offenders' right to mental health seems to outweigh societal interests in their suffering and entitles them to therapeutic memory alterations.

A Duty to Enhance

Finally, Vedder and Klaming have proposed employing memory enhancements for the common good, e.g., to improve recollection of witnesses (Vedder and Klaming 2010). In their view, public interests in obtaining accurate testimony might outweigh freedom of memory of witnesses. Although for lack of reliable enhancers such calls are premature at present, it is worth pondering whether states could stipulate duties to enhance memory. Freedom of memory does not oppose voluntary but protects against mandatory enhancements. As argued before, duties of witnesses seem to end where (more than trivial) side effects begin. But even for apparently side-effect-free memory enhancements like regular sleep (Thorley 2013). or repeated retrieval of the respective memory (Chan and Lapaglia 2011; Pansky and Nemets 2012), a legal obligation might be considered an undue burden to witnesses. Further details depend on particularities of each jurisdiction. Some do not even require witnesses to "refresh" their memory, e.g., through consulting records, while others explicitly ban memory-altering tools, including enhancements.[37] If lawmakers are tempted to amend current provisions, more than side effects should be taken into account. Vedder and Klaming's proposal is intriguing because it prompts general considerations over the politics of memory.

Conclusion and Future Directions

For the most part of history, humankind's knowledge of the past has relied on orally transmitted accounts based on memories, individual and collective. Memory has been an extremely important capacity for cultural development, reflected in the

[36] Some countries have mandatory rehabilitation programs for drug users and sex offenders.
[37] Cf. § 136a German Code of Criminal Procedure.

fact that duties to remember can be found in the writings of many major religions (Margalit 2002). Forgetting has likewise been a tool for political power. Imperatives to erase names of disgraced persons can be found in the bible. A Roman punishment consisted in *damnatio memoriae*, the "condemnation of the memory" of a person. These were predecessors of notorious attempts to manipulate collective memory such as Stalinist retouching of historical photographs. Ricoeur points to the Edicts of Nantes by Henry IV whose very first article proclaimed that "the memory of all things that have taken place" during preceding wars "remains extinguished and dormant as something that has not occurred" (Ricoeur 2004, p. 454). Back then, amnesia was prescribed to secure peace. Today we can witness converse tendencies such as granting amnesty for people who truthfully testify before South Africa's Truth and Reconciliation Committee. Both examples prove the political dimension of memory.

Calls to employ enhancements for public interests should be viewed with suspicion. They would, for the first time, expand the reach of governmental powers over memory from external records directly into minds and brains. What if states enhance only those memories supportive of their cause – might the past not be alterable through partially enhancing recollection? What if other memories were subdued or erased? How could misuse of these powers be effectively ruled out?

The inaccessible nature of one's own and others' memories has shielded them from direct governmental access. Neuroscience might overcome these natural barriers with potentially far-ranging implications for the way society deals with memories of individuals and with history at large. They need to be discussed not only by ethicists but the general public and democratic institutions, in order to formulate a framework for a politics of memory. On occasion, the memories of a single individual have changed the course of history. This perspective might underscore the importance of freedom of memory and motivate its acceptance as a fundamental right. After all, as George Orwell urged, if governments acquire the means to alter the way we perceive the past, they may well acquire those to alter the future.

Acknowledgements This chapter is dedicated to the memory of the first author's grandmother, Margarethe Bublitz, who passed away during its writing. The work was funded by a grant from the Volkswagen Foundation, Germany.

Cross-References

▶ Impact of Brain Interventions on Personal Identity
▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
▶ Neuroenhancement
▶ Neuroethics and Identity
▶ Reflections on Neuroenhancement
▶ The Morality of Moral Neuroenhancement

References

Alberini, C. M., Milekic, M. H., & Tronel, S. (2006). Mechanisms of memory stabilization and destabilization. *Cellular and Molecular Life Sciences, 63*, 999–1008.

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410*, 366–369.

Bell, J. (2008). Propranolol, post-traumatic stress disorder and narrative identity. *Journal of Medical Ethics, 34*, e23.

Beracochea, D. (2006). Anterograde and retrograde effects of benzodiazepines on memory. *Scientific World Journal, 6*, 1460–1465.

Bernecker, S. (2010). *Memory: A philosophical study*. Oxford: Oxford University Press.

Blustein, J. (2008). *The moral demands of memory*. Cambridge/New York: Cambridge University Press.

Borges, J. L., Yates, D. A., & Irby, J. E. (2007). *Labyrinths: Selected stories & other writings*. New York: New Directions.

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory* (Oxford psychology series, Vol. 38). New York: Oxford University Press.

Bublitz, J. C., & Merkel, R. (2009). Autonomy and authenticity of enhanced personality traits. *Bioethics, 23*(6), 360–374.

Bublitz, J. C., & Merkel, R. (2014). Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy, 8*(1), 51–77.

Cahill, L., Prins, B., Weber, M., & McGaugh, J. L. (1994). Beta-adrenergic activation and memory for emotional events. *Nature, 371*(6499), 702–704.

Calhoun, L. G., & Tedeschi, R. G. (2006). *Handbook of posttraumatic growth: Research and practice*. Mahwah: Lawrence Erlbaum Associates.

Chan, J. C., & Lapaglia, J. A. (2011). The dark side of testing memory: Repeated retrieval can enhance eyewitness suggestibility. *Journal of Experimental Psychology. Applied, 17*, 418–432.

Coffman, B. A., Clark, V. P., & Parasuraman, R. (2014). Battery powered thought: Enhancement of attention, learning, and memory in healthy adults using transcranial direct current stimulation. *Neuroimage, 85*, Part 3, 895–908.

Conway, M. (2005). Memory and the self. *Journal of Memory and Language, 53*, 594–628.

Corkin, S. (2002). What's new with the amnesic patient H.M.? *Nature Reviews. Neuroscience, 3*, 153–160.

DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge: Cambridge University Press.

Donovan, E. (2010). Propranolol use in the prevention and treatment of posttraumatic stress disorder in military veterans: Forgetting therapy revisited. *Perspectives in Biology and Medicine, 53*(1), 61–74.

Dresler, M., Sandberg, A., Ohla, K., Bublitz, C., Trenado, C., Mroczko-Wąsowicz, A., Kühn, S., & Repantis, D. (2012). Non-pharmacological cognitive enhancement. *Neuropharmacology, 64*, 529–543.

Dudai, Y. (1996). Consolidation: Fragility on the road to the engram. *Neuron, 17*, 367–370.

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience, 35*(1), 227–247.

Dufner, A. (2013). Should the late stage demented be punished for past crimes? *Criminal Law and Philosophy, 7*(1), 137–150.

Eco, U. (1998). An ars oblivionalis? Forget it. *Publications of the Modem Language Association, 103*, 254–261.

Erdelyi, M. H. (2006). The unified theory of repression. *Behavioral and Brain Sciences, 29*, 499–551.

Erler, A. (2011). Does memory modification threaten our authenticity? *Neuroethics, 4*(3), 235–249.

Feinberg, J. (1986). *The moral limits of the criminal law*. New York: Oxford University Press.

Galert, T., & Hartmann, D. (2007). Person, personal identity, and personality. In R. Merkel et al. (Eds.), *Intervening in the brain. Changing psyche and society*. Berlin/New York: Springer.

Genzel, L., Kroes, M. C., Dresler, M., & Battaglia, F. P. (2014). Light sleep versus slow wave sleep in memory consolidation: a question of global versus local processes? *Trends in Neurosciences, 37*, 10–19.

Geraerts, E., & McNally, R. J. (2008). Forgetting unwanted memories: Directed forgetting and thought suppression methods. *Acta Psychologica, 127*, 614–622.

Glannon, W. (2006). Psychopharmacology and memory. *Journal of Medical Ethics, 32*(2), 74–78.

Glannon, W. (2010). The neuroethics of memory. In S. Nalbantian et al. (Eds.), *The memory process, neuroscientific and humanistic perspectives* (pp. 233–251). Cambridge, MA: MIT Press.

Hammer, L. M. (2001). *The international human right to freedom of conscience: Some suggestions for its development and application*. Aldershot/Burlington: Ashgate/Dartmouth.

Harris, C., Sutton, J., & Barnier, A. (2010). Autobiographical forgetting, social forgetting, and situated forgetting: Forgetting in context. In S. Della Sergio (Ed.), *Forgetting* (pp. 253–284). New York: Psychology Press.

Henry, M., Fishman, J. R., & Yougner, S. J. (2007). Propranolol and the prevention of PTSD: Is it wrong to erase the “sting” of bad memories? *American Journal of Bioethics, 7*, 12–20.

Holmes, E. A., Sandberg, A., & Iyadurai, L. (2010). Erasing trauma memories. *British Journal of Psychiatry, 197*, 414.

Hötting, K., & Röder, B. (2013). Beneficial effects of physical exercise on neuroplasticity and cognition. *Neuroscience & Biobehavioral Reviews, 37*(9), 2243–2257.

Hübener, M., & Bonhoeffer, T. (2010). Searching for engrams. *Neuron, 67*(3), 363–371.

Husain, M., & Mehta, M. A. (2011). Cognitive enhancement by drugs in health and disease. *Trends in Cognitive Sciences, 15*, 28–36.

James, W. (2007). *The principles of psychology*. New York: Cosimo Classics.

Joëls, M., Fernandez, G., & Roozendaal, B. (2011). Stress and emotional memory: A matter of timing. *Trends in Cognitive Sciences, 15*, 280–288.

Kandel, E. (2007). *In search of memory: The emergence of a new science of mind*. New York: Norton.

Karpicke, J. D., & Roediger, H. L., 3rd. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Kindt, M., Soeter, M., & Vervliet, B. (2009). Beyond extinction: Erasing human fear responses and preventing the return of fear. *Nature Neuroscience, 12*, 256–258.

Klein, S., German, T., Cosmides, L., & Gabriel, R. (2004). A theory of autobiographical memory: Necessary components and disorders resulting from their loss. *Social Cognition, 22*(5), 460–490.

Kolber, A. (2006). Therapeutic forgetting: Legal and ethical implications of memory dampening. *Vanderbilt Law Review, 59*, 1561.

Kolber, A. (2008). Freedom of memory today. *Neuroethics, 1*(2), 145–148.

Kroes, M. C., Tendolkar, I., van Wingen, G. A., van Waarde, J. A., Strange, B. A., & Fernández, G. (2014). An electroconvulsive therapy procedure impairs reconsolidation of episodic memories in humans. *Nature Neuroscience, 17*, 204–20.

LeDoux, J. (2007). Consolidation: Challenging the traditional view. In H. L. Roediger, III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 171–176). Oxford: Oxford University Press.

Levy, N. (2007). *Neuroethics*. Cambridge/New York: Cambridge University Press.

Liao, S. M., & Sandberg, A. (2008). The normativity of memory modification. *Neuroethics, 1*(2), 85–99. doi:10.1007/s12152-008-9009-5.

Locke, J. (1979). *An essay concerning human understaning*. In P. H. Nidditch (Ed.), Oxford: Clarendon

Loftus, E. (2003). Our changeable memories: Legal and practical implications. *Nature Reviews. Neuroscience, 4*, 231–234.

Loftus, E., & Ketcham, K. (1996). *The myth of repressed memory: False memories and allegations of sexual abuse*. New York: St. Martin’s Griffin.

Lonergan, M., Olivera-Figueroa, L., Pitman, R., & Brunet, A. (2013). Propranolol's effects on the consolidation and reconsolidation of long-term emotional memory in healthy participants: A meta-analysis. *Journal of Psychiatry & Neuroscience, 38*(4), 222–231. doi:10.1503/jpn.120111.

Lynch, G., Palmer, L. C., & Gall, C. M. (2011). The likelihood of cognitive enhancement. *Pharmacology, Biochemistry, and Behavior, 99*, 116–129.

Margalit, A. (2002). *The ethics of memory*. Cambridge, MA: Harvard University Press.

McGaugh, J. L. (2000). Memory – Century of consolidation. *Science, 287*(5451), 248–251.

Morgan, C., Southwick, S., Steffian, G., Hazlett, G. A., & Loftus, E. F. (2013). Misinformation can influence memory for recently experienced, highly stressful events. *International Journal of Law and Psychiatry, 36*(1), 11–17.

Mullins, J. (1996). Has time rewritten every line? Recovered memory therapy and the potential expansion of psychotherapist liability. *Washington Law Review, 53*, 763–802.

Myers, K. M., & Davis, M. (2007). Mechanisms of fear extinction. *Molecular Psychiatry, 12*, 120–150.

Nadel, L., & Sinnott-Armstrong, W. (Eds.). (2012). *Oxford series in neuroscience, law, and philosophy. Memory and law*. Oxford/New York: Oxford University Press.

Nader, K., & Einarsson, E. O. (2010). Memory reconsolidation: An update. *Annals of the New York Academy of Sciences, 1191*(1), 27–41.

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature, 406*(6797), 722–726.

Nehlig, A. J. (2010). Is caffeine a cognitive enhancer? *Alzheimer's Disease, 20*(Suppl 1), S85–S94.

Pansky, A., & Nemets, E. (2012). Enhancing the quantity and accuracy of eyewitness memory via initial memory testing. *Journal of Applied Research in Memory and Cognition, 1*, 2–10.

Parens, E. (2010). The ethics of memory blunting and the narcissism of small differences. *Neuroethics, 3*(2), 99–107.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon.

Parker, E., Cahill, L., & McGaugh, J. (2006). A case of unusual autobiographical remembering. *Neurocase, 12*, 35–49.

Parsons, R. G., & Ressler, K. J. (2013). Implications of memory modulation for post-traumatic stress and fear disorders. *Nature Neuroscience, 16*(2), 146–153.

Pitman, R. K. (1989). Post-traumatic stress disorder, hormones, and memory. *Biological Psychiatry, 26*(3), 221–223.

Pitman, R. K. (2011). Will reconsolidation blockade offer a novel treatment for posttraumatic stress disorder? *Frontiers in Behavioral Neuroscience*, *5*:11.

Pitman, R. K., Sanders, K., Zusman, R., Healy, A., Cheema, F., Lasko, N., et al. (2002). Pilot study of secondary prevention of posttraumatic stress disorder with propranolol. *Biological Psychiatry, 51*, 189–192.

Pitman, R.K., Rasmusson, A. M., Koenen, K. C., Shin, L. M., Orr, S. P., Gilbertson, M. W., Milad, M. R., & Liberzon, I. (2012). Biological studies of post-traumatic stress disorder. *Nature Reviews Neuroscience, 13*(11), 769–787.

Poundja, J., Sanche, S., Tremblay, J., & Brunet, A. (2012). Trauma reactivation under the influence of propranolol. *European Journal of of Psychotraumatology, 3*. doi:10.3402/ejpt.v3i0.1547.

President's Council on Bioethics. (2003). *Beyond therapy: Biotechnology and the pursuit of happiness*. Washington, DC: President's Council on Bioethics.

Ramirez, S., Liu, X., Lin, P. A., Suh, J., Pignatelli, M., Redondo, R. L., Ryan, T. J., & Tonegawa, S. (2013). Creating a false memory in the hippocampus. *Science, 341*(6144), 387–391.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews, 93*, 681–766.

Repantis, D., Laisney, O., & Heuser, I. (2010a). Acetylcholinesterase inhibitors and memantine for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 61*, 473–481.

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010b). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*, 187–206.

Ricoeur, P. (2004). *Memory, history, forgetting*. Chicago: University of Chicago Press.

Roediger III, H. L., Weinstein, Y.,& Agarwal, P. (2010). Forgetting: Preliminary considerations. In S. Della Sergio (Ed.), *Forgetting* (pp. 1–22). New York: Psychology Press.

Roig, M., Nordbrandt, S., Geertsen, S. S., & Nielsen, J. B. (2013). The effects of cardiovascular exercise on human memory: A review with meta-analysis. *Neuroscience and Biobehavioral Reviews, 37*, 1645–1666.

Sacktor, T. C. (2008). PKMζ, LTP maintenance, and the dynamic molecular biology of memory storage. *Progress in Brain Research, 169*, 27–40.

Schacter, D. (1995). Memory distortions: History and current status. In D. Schacter (Ed.), *Memory distortions. How minds, brains and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Schacter, D. L. (2002). *The seven sins of memory: How the mind forgets and remembers* (1st ed.). Boston: Houghton Mifflin.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 362*(1481), 773–786.

Schacter, D. L., & Loftus, E. F. (2013). Memory and law: What can cognitive neuroscience contribute? *Nature Neuroscience, 16*(2), 119–123.

Schacter, D. L., Addis, D., & Buckner, R. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews. Neuroscience, 8*, 657–661.

Schechtman, M. (2005). Personal identity and the past. *Philosophy, Psychiatry, & Psychology, 12*(1), 9–22. doi:10.1353/ppp.2005.0032.

Schiller, D., & Phelps, E. A. (2011). Does reconsolidation occur in humans? *Frontiers in Behavioral Neuroscience*, *5*. doi:10.3389/fnbeh.2011.00024.

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*(7277), 49–53.

Schudson, M. (1995). Dynamics and distortions in collective memory. In D. Schacter (Ed.), *Memory distortions. How minds, brains and societies reconstruct the past* (pp. 346–364). Cambridge, MA: Harvard University Press.

Shema, R., Sacktor, T. C., & Dudai, Y. (2007). Rapid erasure of long-term memory associations in the cortex by an inhibitor of PKM. *Science, 317*(5840), 951–953.

Shen, F. (2012). Monetizing memory science: Neuroscience and the future of PTSD litigation. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Oxford series in neuroscience, law, and philosophy. Memory and law* (pp. 325–357). Oxford/New York: Oxford University Press.

Smith, M. A., Riby, L. M., Eekelen, J. A., & Foster, J. K. (2011). Glucose enhancement of human memory: A comprehensive research review of the glucose memory facilitation effect. *Neuroscience and Biobehavioral Reviews, 35*, 770–783.

Snead, C. (2011). Memory and punishment. *Vanderbilt Law Review, 64*, 1195–1264.

Strange, B. A., Kroes, M. C., Fan, J. E., & Dolan, R. J. (2010). Emotion causes targeted forgetting of established memories. *Frontiers in Behavioral Neuroscience, 4*, 175.

Strawson, G. (2011). *Locke on personal identity: Consciousness and concernment* (Princeton monographs in philosophy). Princeton: Princeton University Press.

Suthana, N., & Fried, I. (2014). Deep brain stimulation for enhancement of learning and memory. *Neuroimage, 85*, Part 3, 996–1002.

Sutton, J. (2012). Memory. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/memory/.

The British Psychological Society. (2008). Guidelines on memory and the law: A report from the Research Board, Leicester. June 2008.

Thompson, R. F. (2005). In search of memory traces. *Annual Review of Psychology, 56*(1), 1–23.

Thorley, C. (2013). The effects of recent sleep duration, sleep quality, and current sleepiness on eyewitness memory. *Applied Cognitive Psychology, 27*(5), 690–695.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*(1), 1–12.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review Psychology, 53*, 1–25.

Vaiva, G., Ducrocq, F., Jezequel, K., Averland, B., Lestavel, P., Brunet, A., & Marmar, C. R. (2003). Immediate treatment with propranolol decreases posttraumatic stress disorder two months after trauma. *Biological Psychiatry, 54*(9), 947–949.

Vedder, A., & Klaming, L. (2010). Human enhancement for the common good – Using neurotechnologies to improve eyewitness memory. *AJOB Neuroscience, 1*(3), 22–33.

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: State of the art. *Annual Review of Clinical Psychology, 9*, 215–248.

Walker, M. P., & van der Helm, E. (2009). Overnight therapy? The role of sleep in emotional brain processing. *Psychological Bulletin, 135*, 731–748.

Wilson, A. E., & Ross, M. (2001). From chump to champ: People's appraisals of their earlier and present selves. *Journal of Personality and Social Psychology, 80*(4), 572–584.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235–269.

Worthen, J. B., & Hunt, R. R. (2010). *Mnemonology*. New York: Psychology.

Cognitive Liberty or the International Human Right to Freedom of Thought

83

Christoph Bublitz

Contents

Abstract

The aim of the chapter is to draw attention to a fundamental right that is neglected in the law but that is highly relevant for neuroethics: cognitive liberty or freedom of thought. Although an internationally accepted human right, it has not gained practical legal importance. However, any regulation of neurotechnologies has to be evaluated in its light. As the right is unfamiliar to policy makers and even to many lawyers and legal scholars, its historical development and the main arguments for its recognition are sketched. Furthermore, some suggestions for its

C. Bublitz
Faculty of Law, University of Hamburg, Hamburg, Germany
e-mail: christoph.bublitz@uni-hamburg.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_166,

interpretation, scope, and contours are forwarded and remaining open questions identified. According to international human rights law, the right is of absolute nature so that interferences cannot be justified for interests of the common good or paternalistic reasons. Whether this strict prohibition of intervening into other persons' minds can and should be sustained even in light of putative pressing public interests and various neuroethical considerations is one of the novel questions neuroscience poses for the law.

Introduction

"Thought is free" ranks among the prominent tenets of western philosophical and political thought, found in writings and speeches from Cicero and Shakespeare to solemn political proclamations. Freedom of thought is a big, influential, and, at the same time, an amorphous idea. It is not always evident what people mean when they allude to it. For one, it can describe a factual property of psychological entities called thoughts. "Thought is free" is then a statement that can be true or false, depending on the conditions of freedom and their realization, with freedom mainly pertaining to the independence of thoughts from constraints or determining causes. In a different sense, freedom of thought designates a normative claim, thoughts *shall* be free. As a right, it imposes duties on persons, primarily to refrain from interferences with thoughts and thinking processes of someone else. Freedom, in this normative sense, means that others do not have legal claims over thoughts of another person. The manifold statements praising freedom of thought do not often keep factual and normative meanings separate. At times, freedom of thought is even used interchangeably with nearby concepts such as freedom of speech. Of present interest is the normative sense: cognitive liberty as a political demand for liberty of thought. Still, it is worth pondering briefly over the first meaning too. What does the view that thought is free, widely held in commonsense psychology, mean?

Most of the time, we experience our thoughts as free, for instance, when we rather passively observe thoughts popping up and vanishing in our stream of consciousness. This phenomenological freedom seems to result from the absence of any felt causes or constraints of thoughts. Moreover, thinking is an activity, comprising a bundle of diverse mental capacities over which we have powers and conscious control to varying degrees. We possess some powers to direct our thoughts, e.g., when we deliberately set out to think about an issue or use our imagination. Although we do not normally think which thoughts we will entertain next – which would itself constitute a thought and would lead into an infinite regress – we can roughly steer the contents of subsequent thoughts and thus seem to think what we want and to be in control of our thoughts. At other times, by contrast, we cannot fully grasp our thoughts, suppress unwanted thoughts from circling through our heads, or stop the wandering of our mind. In those cases, control over thoughts is limited; hence, they might appear as "unfree." Apart from those and other (pathological) cases, the feeling of freedom of thought is an inescapable aspect of our everyday conscious experience.

In another sense, thoughts are free because of their private character. No one else knows the content of thoughts in the same way as the thinker does. Thoughts are not directly observable for others; they can only be inferred from verbal or behavioral expressions. In addition to this privileged epistemic access that confers authority over the knowledge of one's thoughts, privacy of thoughts can also mean that others cannot control our thoughts because they are inaccessible from the outside. It seems impossible to compel another person to entertain a particular thought or to implant ideas or opinions. One can speak to another person, irritate, annoy, or bombard her with all kinds of stimuli, but such interventions are so imprecise and indirect that they can be blocked by inner countermeasures in all but the most extreme cases. These characteristics of thought seem to roughly capture what is commonly meant by freedom of thought. They are nicely condensed in the lines of an old German folk song: *No scholar can map them/No hunter can trap them/No man can deny*: *Die Gedanken sind frei*! (*Thoughts are free*).

These properties of thoughts loosely map onto two forms of freedom that are typically distinguished: internal freedom as independence from inner causes or influencing factors such as emotions and external freedom as the absence of external forces that can monitor or alter thoughts. Some of the apparently familiar characteristics of thought just mentioned raise deep philosophical questions: In which way do we have privileged access to our thoughts – is thinking something like inner perception of mental objects? Can we sometimes err about our thoughts? Might not the external world determine the content of our thoughts? What is the relation between thought and thinker? And what are thoughts anyway? These questions already indicate that our intuitive idea of freedom of thought is in need of further exploration.

The picture of thoughts and the mind as an invincible realm of inner freedom, an inner citadel to which the world has no entrance, is in many ways questionable and misleading. It is deeply rooted in a Cartesian dualism of mind and matter. However, since the day's of Freud, psychology assumes to various degrees that thoughts are generated by a complex and largely unconscious psyche which apparently follows its own rules and dynamics. For instance, thoughts do not line up randomly but seem to stand in some relation to one another, yet we are often unable to introspectively identify the corresponding rule. Today, neuroscience approaches the workings of the mind via its material substrates on various levels, from the connectivity of brain networks down to molecular mechanisms. Irrespective of whether or not reductionist strategies will ever succeed in providing a complete explanation of mental phenomena including thoughts, the initially appealing idea of thoughts being independent from internal causes and influences cannot but appear as a naïve folk psychological belief, based on the transparent bounds of introspection that conceal the operations of the mind and brain from the inward gaze.

Furthermore, many psychological processes are not under direct conscious control. We cannot, e.g., alter our mood at will. Some types of thoughts, such as beliefs and opinions, are not free in the sense that we can bring them about

voluntarily because they adhere to rules of logic or semantics. And of course, many of our desires are not under our voluntary control either. Even those mental capacities that are consciously controllable such as reasoning, remembering, or calculating are limited nonetheless as they require depletable mental resources. Thus, although we have varying degrees of conscious control over our minds, including our thoughts, our internal freedom is limited in many ways.

Likewise, we may not enjoy a strong form of external freedom. Our thoughts are not sealed off from the outside world. Although thought control in the strict sense of directly controlling the content of another's thoughts is not (yet) possible, there are many indirect ways, from classic conditioning to psychoanalysis and cognitive therapy, that may can effectively change the way people think. The neurobiology of the brain has been targeted by more direct means since the days of crude forms of psychosurgery in the 1930s which proved that the mind is not out of reach of external forces.[1] At least since the pharmacological revolution of the 1950s, interventions aiming at altering mental states, capacities, and thoughts are widely employed by the state and psychiatry, not always based on consent of affected persons. While pharmaceuticals allow modulating some properties of thought such as speed or style of thinking, novel interventions afford targeting thoughts more specifically. Deep brain stimulation (DBS) may give a glimpse of the near future. In studies, stimulation altered recurrent obsessive-compulsive thoughts as well as thoughts related to depression and anorexia nervosa (Lipsman et al. 2013; Schlaepfer et al. 2007). Moreover, if the hypothesis of "concept cells," neurons that respond to particular concepts or ideas, proves correct, stimulation of particular neurons may enable to elicit specific thoughts (Quiroga 2012). Even though thought control is still far off, it seems that the degree to which thoughts can be accessed or altered through interventions on the biological level is merely a contingent matter.

Freedom of Thought

The advent of tools that confer more powers over our own and other's mental realm leads to the normative sense of freedom of thought. The question no longer is whether it is possible to change thoughts, but rather who should be allowed to do so, by which means, and for which purposes. In order to regulate novel mind-altering tools and to evaluate existing regulations, the law has to confront a fundamental question: What is the legal relation of a person to her mind? Which legal (and enforceable) claims do others, including the state, have over the mind of someone else? The topic is rarely addressed in this abstract way. Sometimes, it is couched in terms of "ownership" of minds, but in a strict sense, one cannot "own" one's mind in the same ways one can have – and sell or relinquish – property in external things.

[1]For the history of psychosurgery and electrical brain stimulation, see Valenstein 1974; Delgado 1969.

The relationship between a person and her mind is qualitatively different. After all, if our very existence is, as one may understand Descartes' "cogito" argument, predicated upon our thinking, what could be more constitutive of a person than her mind and mental states?

Advocates of cognitive liberty demand that the individual should enjoy a wide range of autonomy over what is on – and in – her mind. The term was coined by civil right activists of the Center for Cognitive Liberty and Ethics (CCLE). Their central vision has been elucidated in a series of programmatic papers by Richard Boire and Wrye Sententia (Boire 1999, 2000, 2001, 2005; Sententia 2004), in short: To secure "the right to control one's own consciousness [as] the quintessence of freedom."

Historical Glance

In the last decade, "cognitive liberty" has become a political slogan in support of various related causes such as the decriminalization of psychoactive substances or restrictions of involuntary interferences with other minds, from advertisement to coerced psychiatry. The underlying idea, however, is anything but new. Abstractly, it is best understood as the postulate that other persons – or the state – do not have claims over the contents of other persons' minds. Variations can be traced throughout the history of ideas. An early formulation is a maxim of Roman law: *cogitationis poenam nemo patitur*, no one shall be punished for his thoughts alone. It restricted punishable offences to outward behavior and liberated thoughts and opinions from the sovereign's control to some extent. In Christian theology, the freedom to (not) believe in the truth of scripture and the question whether force can and should be used to compel disbelievers to change purportedly errant beliefs were a recurrent theme since Augustine. In medieval times, competing ecclesial and worldly authorities delineated their jurisdictions and purviews of power such that the church took care of saving souls (so-called forum internum), whereas worldly authorities regulated external actions (forum externum). Thereby, the mind was placed outside the reach of worldly powers. It was one of the major goals (and achievements) of the Reformation to free the forum internum from ecclesial authority by proclaiming a direct connection between the conscience of man and God. Because God was held to rule directly through the individual's conscience, the mediating role of the church could be circumvented. The acceptance of freedom of belief and conscience and its exemption from access by authorities was a prerequisite of toleration, the basis to end religious strifes that rippled Europe for centuries. Historically, freedom of belief and conscience are conceived as the first human rights.[2]

Freedom of thought is, of course, the overarching motto of the Enlightenment, expanding freedom of belief to opinions on nonreligious matters. Kant famously

[2]For the history of freedom of thought up to the early twentieth century, see Bury 1952.

answered the question "what is enlightenment" with the claim that everyone ought to dare thinking for themselves. The demand of freedom of thought was a key element in the political struggles that paved the way for modern western democracies, and it was understood broadly as the uncensored exchange of ideas, freedom from persecutions for holding and voicing opinions, and freedom of expression and the press. Today, these freedoms are firmly established and well-defined guarantees of the human rights system and are protected by freedom of speech and expression which often entail the freedom to "hold opinions without interferences."[3]

Freedom of thought as such, however, has never received much attention although the mind has always been a central point of political, ideological, and social struggles. States always had an interest in mentally disciplining citizens. As Spinoza wrote, those rulers "exert the greatest power who reign in the hearts and minds of their subjects" and "were it as easy to control people's minds as to restrain their tongues, every sovereign would rule securely" (Spinoza 2007, p. 208). One of the first horrific examples of the extent to which the human mind can be forcefully bent was provided by Maoist thought-reform programs that successfully altered people's opinions through a mixture of deprivation, indoctrination, and sophisticated psychological means (Lifton 1989; Taylor 2006). The west responded to this communist "brainwashing" with research efforts, among others the infamous clandestine US MKUltra program which conducted experiments with mind-altering techniques, sometimes on unknowing subjects (Marks 1992; for behavioral modification Macklin 1981). Aware of the importance of neurobiology, psychologist William Sargant wrote about the ideological battle during the cold war that "the political-religious struggle for the mind of man may well be won by whoever becomes most conversant with the normal and abnormal functions of the brain and is readiest to make use of the knowledge gained" (Sargant 1997).

The International Human Right to Freedom of Thought

The importance of the mind has been reflected in human rights law. The freedom of the inner realm has been expanded from religious beliefs and conscience to thoughts in general. Article 18 of the Universal Declaration of Human Rights, adopted in 1948, explicitly guarantees that "everyone has the right to freedom of thought, conscience and religion." In an almost identical wording, it has been incorporated into almost every human rights treaty[4] and is thus a binding – and in

[3]Art. 19 Universal Declaration of Human Rights (UDHR); Art. 19-1 International Covenant on Civil and Political Rights (CCPR).

[4]Art. 18-1 CCRP; Art. 9-1 European Convention on Human Rights (ECHR); Art. 13 American Convention on Human Rights.

theory, effective – universal human right.[5] Many national constitutions, however, do not explicitly recognize such a right on the domestic level.[6] Reference to freedom of thought is nonetheless made sometimes in decisions of high courts. The US Supreme Court, for instance, held that freedom of thought "is the matrix, the indispensable condition, of nearly every other form of freedom. With rare aberrations a pervasive recognition of this truth can be traced in our history, political and legal."[7] These words sound as if freedom of thought was understood as a general principle of law. However, it is often not mentioned in cases prima facie affected by it. As an effective legal right, the appealing yet vague idea needs to be rendered more precise.

What does being entitled to think freely mean? Courts have never defined content and scope of the right in more detail. Even legal scholarship only provides some approximations. According to (nonbinding) commentaries of human rights treaties, it affords the "freedom to entertain any thought [or] moral conviction."[8] In combination with conscience and belief, freedom of thought "cover[s] all possible attitudes of the individual towards world and society"; it protects the "absolute character of the freedom of an inner state of mind"[9] and the liberty to "develop autonomously thoughts and conscience free from impermissible external influence" (Nowak 2005, p. 412). Accordingly, it prohibits being "subjected to treatment intended to change the process of thinking," "any form of compulsion to express thoughts [or] to change an opinion" (Vermeulen, supra, p. 752), or influence "of the conscious or subconscious mind with psychoactive drugs or other means of manipulation" (Nowak 2005, p. 413). In short, freedom of thought provides protection against severe interventions into minds that aim at altering thoughts or thinking processes and thereby opposes the use of most novel neurotechnologies on non-consenting persons.

Some identify a more abstract principle underlying the right, a general restriction of state power in line with the right's historical genesis: "The inner world of the person lies outside the jurisdiction of the state" (Harris et al. 2009, p. 428). In other words, the state has, in principle, no claims over minds of citizens. Yet again, this principle is neither widely accepted nor openly contested in contemporary constitutional theory. The relation of governments to the minds of citizens is an underexplored issue at present. At least, freedom of thought seems to be the reason for an interesting feature of positive law: The law is generally reluctant to directly

[5] Cf. the General Comment No. 22 30.07.1993 by the UN Human Rights Committee.

[6] This touches upon the problem of the interplay between international human rights and domestic law. Whether and to which extent human rights are enforceable on the domestic level differs from treaty to treaty and country to country. In some states, human rights treaties prevail over domestic law; in others, they have to be incorporated into domestic law through special legislation; in still others, domestic law has to be interpreted in light of international treaties.

[7] Palko v. Connecticut 302 U.S. 319 (1937). Cf. Blitz (2010) and Boire (1999) for further references in US jurisprudence.

[8] Vermeulen in van Dijk and van Hoof 2006, p. 752.

[9] Scheinin in Eide and Swinehart 1992, pp. 264/266.

regulate minds. Only few, if any, legal provisions explicitly stipulate that persons ought (not) to be in a specific mind state. Instead of minds, the law regulates behavior by pre- or proscribing external actions.

Practical Irrelevance

While courts and scholars have often reiterated the importance of the right, it has not made much difference in legal proceedings. Freedom of thought may very well be the only human right without any real practical application. One can only speculate about the reasons. For one, the law is generally hesitant to deal with ostensibly intangible mental states, mainly because of issues of proof and causation. Roughly, the law entertains a dualistic protection of the person. While bodies enjoy strong protection and every constitution protects bodily integrity, one only rarely finds similar provisions for mental integrity. Compensation for mental distress is only awarded in extreme cases, many jurisdictions do not even recognize, e.g., a tort covering mental harms without corresponding physical injury.

Moreover, legal thinking appears still permeated by folk psychological intuitions and impressed by the view that thoughts are by their very nature invincible. In 1942, the US Supreme Court was convinced that "[f]reedom to think is absolute of its own nature; the most tyrannical government is powerless to control the inward workings of the mind."[10] It seems that not much has changed. The law might not have the manifold effective ways to change minds in plain view, and this could partly be due to the way in which worrisome interventions are described. Terms such as "thought control" or "mind-control" evoke sinister images of universally ostracized actions like brainwashing, but may not adequately capture, e.g., injections of mind-altering substances. Many interventions are imprecise in their effects so that it is hard to understand them as exerting "control" over thoughts.

Absolute Protection

The practical irrelevance of the right has another important reason: In most human rights treaties, freedom of thought enjoys *absolute* protection, for which interferences cannot be justified by any countervailing interest. Unlike regular rights which have to be balanced against rights of others or public interests, absolute rights trump other considerations. Human rights law recognizes only few absolute rights (e.g., prohibition of torture or degrading treatment) which evince the exalted standing of freedom of thought.[11] This importance, however, has a downside in practice which can be illustrated by coerced psychiatry. Obviously, forcible administration of

[10]Jones v. Opelika, 316 U.S. 584 (1942).

[11]Under the UDHR and the ICCPR, the right is even "non-derogable" in times of emergency.

psychoactive substances modulating thoughts and emotions is the kind of treatment that the right prohibits. Because of its absolute nature, interventions are not justifiable on any ground. In practice, however, freedom of thought is not even mentioned or elaborated upon in respective judgments. Instead, it is debated whether patients have a (restrictable) "right to be mentally ill" or "to refuse treatment." Treatments are mainly considered on the bodily level (side effects of medications), which puts the genuine mental effects out of focus. Thus, in face of practical necessities and arguably ethical duties to interfere with the inner realm of citizens in some cases, freedom of thought is simply not applied or construed so narrowly that measures are not conceived as infringements. Narrowing the scope of rights to avoid having to justify interferences is an old conjuring trick of lawyers, with the paradoxical effect that those rights that should command utmost respect are belittled and stripped of practical importance. This is not to contend that coerced treatments necessarily violate human rights, as some civil rights groups maintain. But one would expect sound arguments in view of such evident interferences, and their absence indicates that the absolute protection of the mind is an idea that stems from a time in which means to coercively alter thoughts were not available (or even conceivable) and that encounters deep ethical concerns once they become so, as there might be alluring reasons to employ them. As an alternative to degrading the right to a proclamation of symbolic nature, it seems preferable to cast doubts on the absolute protection and argue for well-defined exceptions while maintaining the strong protection in principle.

Altering One's Own Mind

As formulated today, cognitive liberty comprises another dimension: the right to alter one's own mind, not only by one's natural capacities but also with the help of neurotools from pharmaceuticals to brain stimulation. It is important to note that the current law does not outlaw particular mental states. There is no prohibition to be "high" or in any other altered state of consciousness, just as there is none to have criminal thoughts, deviant desires, or enhanced mental skills. However, some means to attain those states are prohibited by drug regulations outlawing the non-licensed use and possession of substances from classic street drugs like marijuana to potential neuroenhancers like methylphenidate. Details of regulations and criminalization vary from one country to the next, but international drug control treaties impose a ban on many substances in nonmedical contexts.[12] The idea that persons may have a legitimate interest in their use for nonmedical purposes does not play a role in current drug regulation. From a legal perspective, the debate over cognitive enhancement is thus not very different from the one about decriminalizing classic

[12]Cf. The list of psychotropic substances under International Control by the International Narcotics Control Board (www.incb.org).

recreational drugs. Both directly pertain to the question to which extent persons should be free to alter their minds for nonmedical purposes of their own liking.

At the moment, attaining some mental states is factually only possible by breaking the law. However, as the law does not prohibit altered mental states, persons have a right to be in those states. This right would be weighty if it derived from freedom of thought, but the point has not been thoroughly addressed so far and depends on the exact construal of the right. Historically, this dimension has not been part of the guarantee of the right. However, as commentators suggest, freedom of thought implies the freedom to "entertain any thought." A strong right to be in mental state X must prima facie entail the permission to use the pathways to arrive at or attain X (as long as interests of others are not affected). Then, freedom of thought entails the permission to use mind-altering substances. Although such arguments have not been accepted by courts (yet), much speaks in their favor.[13]

More generally, it is interesting to see how different ways to alter one's own mind fall under different regulations. Novel tools such as tDCS or TMS are not regulated at all (Maslen et al. 2013). Other mind-changing stimuli enjoy quite strong legal protection, e.g., as part of freedom of speech and expression. In a landmark case concerning pornography, the US Supreme Court even suggested that control of stimuli that enter minds could amount to mind control.[14] At least, compelling reasons are needed to justify why curbing one type of stimuli interferes with precious rights, whereas curbing others does not. In light of freedom of thought, one cannot but suspect an inconsistency here. Different means may warrant different treatment due to safety issues, but the same strong liberty interest for using of mind-altering tools would have to be conceded before one can assess whether it is outweighed by potential risks in particular cases.

After all, at least in theory, a strong human right protects persons against unwanted interventions into their minds. It arguably also entails a prima facie permission to make use of mind-altering tools. This right is, however, of little practical importance. Freedom of thought has failed to stand the test of practical applicability; it did not exert much influence on court cases or drug policy presumably because the idea is too vague and – due to its absolute nature – the right too strong to provide reasonable guidelines for solving practical cases.

Arguments for Cognitive Liberty

This gives rise to the challenge for the law – and for democratic societies – to draw the outer boundaries of the person in her mental aspects. Many substantive issues over scope and content of the right have to be clarified. To this end, some no longer sustainable assumptions, such as the idea that thoughts are factually invincible,

[13]For a moral argument to that end, see Husak (1992).

[14]Stanley v. Georgia – 394 U.S. 557 (1969).

have to be jettisoned. On the normative side, an absolute prohibition of any measure that interferes with thoughts and thinking processes appears unreasonable, at least if the right is, as suggested here, construed widely to encompass the use of mind-altering substances. Likewise, interventions that restore mental capacities in incompetent persons are at least not unreasonable and should not be prohibited across the board a priori. Rather, exceptions to the absolute protection have to be cautiously discussed and defined.

Moreover, some open questions about the scope of the right can be identified: Current law primarily covers some mental entities – thoughts and opinions. Should legal protection be confined to them (whatever they exactly are) or encompass others? Neurointerventions target, e.g., emotions such as pleasure and pain, motivation and anhedonia, mood in general, volition, and other psychological processes and properties that may not self-evidently qualify as thoughts, but which are so intimately intertwined with thoughts – if separable at all – that a strong protection of one without protection of the other does not appear meaningful. As the mind becomes accessible on the cerebral level, the law may need to provide an all encompassing protection – freedom of the mind. On the other hand, the same level of protection for all mental states might not account for their differences. Memory, for instance, seems to be a special case in its significance for both the individual and society.[15]

Historical Arguments

Furthermore, the importance of the right has to be assessed. This requires a deeper theoretical exploration of the reasons for and the limits to mental freedom. Many contemporary calls for employing neuroscience for public goals do not exhibit awareness of the fact that a governmental permission to alter the way people think or feel is, at least in the eyes of the law, a paradigm shift from regulating citizen's conduct to governing their minds. The mind of citizens is not just another target for governmental regulation. It is what makes persons who they are. That is why mind interventions are among the most intimate conceivable invasions of the person.

As the right is not accepted in many national legal systems, it is necessary to sketch some of the reason for its acceptance. A first step is to excavate the historical significance of the right and the distinction between regulating conduct and minds. It might be noteworthy that even in the conception of an (almost) absolutist sovereign in the work that stands at the inception of modern political philosophy, Thomas Hobbes' Leviathan, a distinction between the freedom of inner beliefs and outward confession was drawn. While states can demand public confession of adherence to official doctrine, citizens remain free to believe as they wish in their hearts. State powers are restricted to regulating external

[15]Cf. the chapter on memory alterations and the law in this volume.

actions (Hobbes 1996, p. 306). This idea can be found in the works of many prominent authors. In the Doctrine of Right, Immanuel Kant formulated the idea, concededly oversimplified here, that the role of the state is to secure reciprocal equal freedoms. The freedom of one person can interfere with another person's freedom only if their actions collide in the external world, e.g., because two persons claim access to the same space or resources. By contrast, whatever happens in the interior of a person's mind never restricts the freedom of anyone else. The purview of legitimate legal coercion is thus confined to the regulation of outward actions (Kant 1991, pp. 230, and 214).

In a similar vein, John Stuart Mill emphasized the special role of the mind. In "On Liberty," he wrote:

> [T]here is a sphere of action in which society, as distinguished from the individual, has, if any, only an indirect interest; comprehending all that portion of a person's life and conduct which affects only himself, or, if it also affects others, only with their [. . .] consent. When I say only himself, I mean directly, and in the first instance: for whatever affects himself, may affect others through himself [. . .]. [T]he appropriate region of human liberty . . . comprises, first, *the inward domain of consciousness*; demanding liberty of conscience, in the most comprehensive sense; liberty of thought and feeling; absolute freedom of opinion and sentiment on all subjects [. . .]. Secondly, the principle requires liberty of tastes and pursuits; of framing the plan of our life to suit our own character; of doing as we like, subject to such consequences as may follow; without impediment from our fellow-creatures, so long as what we do does not harm them even though they should think our conduct foolish, perverse, or wrong (Mill et al. 2003, pp. 82–83).

Mill claims that over self-regarding matters, individuals should decide for themselves. The characteristics of a person, her bodily and psychological properties and first of all, the inward domain of consciousness, are the realm which exclusively concern her. Therefore, society does not have claims over those traits, even if it strongly objects to them. Mill's argument relates to his harm principle. Restrictions of freedom are only permissible to avert harm to others. Yet by themselves and apart from deeds, thoughts or opinions cannot harm others. Kant and Mill both dismiss a legitimate right of others or the state over the content of a person's mind, i.e., governments do not have any jurisdiction over minds at all. Consequently, states can only restrict behavior, insofar as it interferes with others' rights or public interests. Of course, one may raise objections: Perhaps the role and corresponding powers of present-day welfare states are broader than securing negative liberties. In addition, thoughts, apart from deeds, may hurt others on occasion.[16] Yet even if one rejects Kantian and Millian ideas, the central point should be acknowledged: It is anything but self-evident why states should have any powers over minds at all.

[16]For example, pedophiliac thoughts as in the US case of Doe v. City of Lafayette, 377 F.3d 757 (2004).

Modern Arguments

Although not often explicitly addressed, the idea of cognitive liberty is deeply entrenched in principles of modern liberal constitutions. Among its defining and widely accepted characteristic is the idea of autonomy. Oftentimes, the scope of autonomy in law is wider than in ethics. Legal permissions do not coincide with what is ethically advisable; they guarantee freedom for supposedly wrong or bad decisions as long as rights of others are not affected. Currently, autonomy is primarily applied to issues regarding a person's body. In many jurisdictions, bodily self-determination reaches so far that even life-saving medical treatments have to be eschewed without consent of the person (or a proxy). Interestingly, one finds much less discussions of "mental autonomy." Of course, the idea of autonomy is not inherently bound to corporeal bodies or the physical part of persons, but rather a general principle of the allocation of decisional competence. The far-ranging autonomy in primarily self-regarding matters granted by the law must equally apply to the mind, including potentially self-harming conduct.

Some further legal arguments should be briefly noted. Together, the diverse human rights form a system of protection with the person and her most intimate characteristics in its center. Many civil liberties and social and political rights expand the protection of the person into the social sphere as they are considered necessary conditions for the development of the personality and a flourishing life. All the other freedoms lose their reference point if the core, the protection of life, liberty, and integrity of the person, is neglected. As the mind is among the most essential aspects of a person, the drafters of the Universal Declaration of Human Rights called freedom of thought "the basis and origin of all other rights" with "metaphysical significance."[17] Boire writes: "If freedom is to mean anything, it must mean that each person has an inviolable right to think for him or herself. It must mean, at a minimum, that each person is free to direct one's own consciousness; one's own underlying mental processes, and one's beliefs, opinions, and worldview. This is self-evident and axiomatic" (Boire 1999).

Freedom of thought stands behind other well-accepted human rights which could be severely undermined without its firm protection (Cf. Blitz 2010). Freedom of speech, for instance, is not merely the right to utter words, but to engage in critical discussions with adequate information. By protecting the uncensored exchange of ideas, the right secures the outer preconditions to form and revise opinions. The deeper reason why states sought and still seek to control information is that they seek to control ideas and opinions. Of course, the mental and neurobiological conditions of forming and revising opinions are at least equally worthy of protection. Because both rights share the same goal, a legal prohibition of governmental censure of ideas without a prohibition of tinkering with brains would be

[17]Rene Cassin, quoted from Scheinin 1992.

inconsistent. Prima facie, the strong protection of freedom of speech has to apply to freedom of thought at least to the same degree.

Moreover, the legal structure and the distribution of rights and responsibilities are modeled upon the idea of a freely deciding person. Apart from larger metaphysical issues of determinism and free will, free decisions presuppose that the preferences on which decisions are made have not been brought about through manipulative influences (above ordinary degrees). The legal system can only legitimately ascribe responsibility and liability for decisions if it simultaneously guarantees a basic level of freedom from unwanted interferences with the decision-making process. If the law treats persons as self-determined and holds them accountable for consequences of their mind states (in criminal and contract law, “meeting of the minds”), it has to grant them the legal powers of self-determination. Then, cognitive liberty is the “right to free will” because it protects the conditions of possibilities of “free” decisions and actions. In light of these considerations, cognitive liberty is not merely a political demand that one can approve or reject, but among the foundations on which liberal constitutional democracies are built. Legal systems seem to presuppose freedom of thought and are bound by reasons of inner consistency to openly recognize and embrace it.

Challenges

Interferences: Do Means Matter?

Once freedom of thought is at least in principle accepted, questions of its violations surface inevitably. As the foregoing shows, the intensity and effects of interventions have to be taken into account. In a sense, we change each other’s minds and brains all the time. Reading these words changes thoughts of the reader. Eric Kandel once remarked that the deeper aim of his lectures is to cause anatomical changes in the brains of his audience. Because minds and brains are constantly changing, often due to external stimuli, the law cannot simply stipulate a prohibition to alter them in the same way as it prohibits altering another person’s body. Rather, the law has to develop a framework of impermissible interventions based on both means and effects of interventions. Interventions that alter opinions, modify decisions, or severely undermine thinking capacities may run afoul of the core guarantee of freedom of thought, whereas minor interventions might not necessarily do so. Boundaries have to be defined more concretely. What about, e.g., subtle shifts in cognitive processes or mood?

Further distinctions have to be drawn between different means to change minds. Whether neurotechnologies are in some way more worrisome than traditional means of altering minds is one of the deep philosophical questions raised by neuroscience. As it has manifold practical implications, it has already reached public discourse and the popular press. It stands in the background of

controversies such as whether children should be treated with methylphenidate or rather placed in better school environments, whether depression should be treated pharmacologically or through psychotherapy, etc. In these cases, the desired goal is often sufficiently clear, but the means to attain them is in dispute.

The issue is so controversial because it is inextricably linked to different worldviews and ideologies. In the 1960s and 1970s, the heyday of behaviorism and leftist movements, the external environment was conceived as the prime determinant of behavior and, e.g., mental disorder. As a consequence, social and economic conditions were conceived as the prime target for interventions. If, as Marxists would say, the social being determines consciousness, consciousness is best changed through changing the social environment. Alternatively, the psychoanalytic tradition identified the psychogenesis of the individual as the main factor and favored various psychoanalytic and psycho-developmental interventions. Then came genetics, and today, we witness a profound shift towards the individual's brain which is considered as potentially maladapted to the environment rather than vice versa. Thus, the brain appears as the most appropriate target of intervention. However, the reductionist view of humans, to some degree inherent to neuroscience, remains unbelievable for many people who thus eschew interventions on the neurobiological level and favor more holistic, spiritual, or natural ones. In short, different levels of explaining persons seem to suggest different means of intervening.

Even though the different explanations might not be as mutually exclusive as they appear at first glance, this metaphysical issue is far from being resolved in the near future. In the meantime, the debate should be freed from ideological parts by separating empirical from more principled questions. Any intervention has benefits and side effects. Whether a desired goal can be better achieved through one way or another is largely an empirical matter. Perhaps, side effects of pharmaceuticals do not outweigh benefits; perhaps, traditional ways have further beneficial effects such as promoting virtues as self-mastery and self-reliance. These are, by and large, empirical matters. The more principled differences remain apart from benefit and risk assessments. To some, neurointerventions are intrinsically worse because their effects take place on the cerebral level whereas others work more indirectly, because they are artificial and technological (as compared to supposedly natural alternatives), or because they alter personality traits and render persons "inauthentic." Such worries are widespread in the concerned public, and corresponding arguments deserve more attention than can be given here. For present purposes, it is necessary to discard crude dualistic assumptions and underline the presupposition that all interventions, from talk therapy to pharmaceuticals or brain stimulation, change the brain in some way. Under this premise, other objections appear as expressions of particular worldviews and thus ill-suited for matters of public policy. As a consequence, a widely held position in neuroethics claims that principled concerns over neurointerventions are unfounded and that means to alter minds are not intrinsically ethically different

and should be treated on par (again, risks and benefits of specific interventions notwithstanding).[18]

In a legal perspective, however, some means to change other minds are privileged. Consider again free speech. It entitles speakers to send particular stimuli, mainly those that qualify as communication because they convey opinions or information. When a speaker changes opinions of a listener, e.g., by a compelling argument, she may have altered thoughts, but not in a way which runs afoul of the idea of freedom of thought. By contrast, if the same effect was produced through a pill that alters opinions, the freedom to hold opinions without interference seems contravened. Individuals have a right against being subjected to unwanted brain stimulation but not necessarily against exposure to persuasive arguments. Consider state interventions with laudable aims. One can hardly deny that governments must possess some competency to influence the minds of citizens, e.g., for purposes of education. Many schools and universities are run by the state, and in the preamble of its constitution, UNESCO declares that "since wars begin in the minds of men, it is in the minds of men that the defences of peace must be constructed." Granting authorities a "teaching power" to instill important social values, such as peacefulness, seems justifiable – but not by any means (Cf. Tussman 1977). Compare a public campaign to foster law-abiding behavior by teaching the categorical imperative and raising awareness of the suffering caused by crime to a more advanced version of a recent study in which social norm compliance was improved by harmless electrostimulation of the lateral prefrontal cortex through transcranial direct current stimulation (tDCS). The authors conclude that their findings of enhanced "voluntary and sanction induced social norm compliance may be of relevance because noncompliance with social norms constitutes a major problem" (Ruff et al. 2013, p. 484). Assume for the sake of argument that main as well as side effects of interventions are relevantly similar – both change opinions and behavior in a pro-social way – should they be treated on par? If tDCS interventions were much cheaper, ceteris paribus, could or should the state require persons to wear stimulating devices instead of putting more effort in public awareness campaigns? Somewhere between these two approaches to promote norm compliance, the difference between Brave New World and a free – but morally conscientious – society seems to be situated.

Although many objections against neurotechnologies may turn out to be unsustainable, some normatively relevant differences between interventions remain. Among the most salient ones in above examples is that persons have more control over particular incoming stimuli such as words compared to electrical stimulation of the brain. The listener can think about propositions, rehearse, revise, or reject them; they enable her to form opinions. Even if the argument is so persuasive that it cannot but lead to a change of opinion, at least from the standpoint of a rational person, no harm has occurred since opinions and beliefs have to stand the test of counterargument. Abstractly, persons have more control over stimuli that

[18]Cf. the different version of the Parity Principle in Levy 2007.

are sensually perceived and enter conscious awareness because it allows them to engage with the contents. By contrast, stimuli that directly change brain activity on other routes and without (or only subsequently) evoking conscious phenomena are less controllable. Whether they are successful solely depends on their strength and biochemical properties of the brain. Stimuli that enter brains via our outward senses are processed by mechanisms whose function it is, by all their faults, to deal with them. To be effective, these stimuli have to engage with the other as a person and have to resonate with her personality, wishes, desires, and psychological properties. Brain stimulation or pharmaceuticals take different routes that circumvent these mechanisms and seem to treat the other more as a natural object. Whereas sensually perceived stimuli are regularly best considered as inputs into the cognitive machinery, direct interventions alter the machinery itself. Again, the difference does not lie in the false assumption that only some interventions cause cerebral changes, but rather in the fact that stimuli enter the brain via different routes and are processed by different mechanisms.

Are these differences normatively relevant?[19] This depends on the kind of duties persons have with respect to minds. If freedom of thought demands that we respect each other's mental sovereignty and forbids manipulating thinking processes, those stimuli that are informational inputs into the thinking machinery and appeal to reason fare much better than those that bypass control mechanism and alter the machinery more directly. Direct stimulation of the brain and persuasive arguments might be conceived as two poles in a broad spectrum of gradually different interventions which have to be carefully analyzed in light of a more fine-grained normative framework. As a first approximation, a concededly rough line can be drawn between interventions that intentionally bypass control capacities or exploit cognitive weaknesses on the one side and interventions that, at least in principle, respect control and freedom of thought as they do not undermine powers of resistance on the other. One example is subliminal messages entering minds through the senses without rising to conscious awareness. Because they are designed to bypass conscious control, they regularly do not respect freedom of thought of the receiver. Within the grey area that needs further analysis lie supraliminal stimuli which aim to change opinions by primarily appealing to emotions or other unconscious dispositions for which neuromarketing provides many examples. Traditional limits to marketing such as "misleading information" may not capture those subtle manipulative interferences which nonetheless appear worrisome in light of freedom of thought. Furthermore, external background conditions can be modified by placing stimuli such as scent, music, or odorless substances as oxytocin in locations like supermarkets to alter customer's behavior, e.g., raising their willingness to make purchases through increasing mood or lowering self-control. How society should deal with such lower-level, perhaps harmless yet potentially effective manipulations is a question that democratic lawmakers have to decide and which requires an open public debate informed by empirical findings.

---

[19] A more detailed argument against the Parity Principle can be found in Bublitz/Merkel 2014.

Paternalistic Limits: For the Good of the Person

Even if interferences are identified, the main challenge lies still ahead. Which interferences can be justified? In its current absolute interpretation, freedom of thought does not allow for any interference. Yet this understanding runs contrary to widespread practice and might be in need of reconsideration. Perhaps, narrowly defined exceptions are unavoidable. Rights are mainly limited for three reasons: for the good of the person herself, in the name of public interests, or because of rights of others. Whether and to what extent freedoms can be curbed for paternalistic reason, i.e., against the will of affected persons but for their own good, is a highly controversial question to which various legal systems provide different answers and which cannot be pursued in detail here. However, coerced psychiatric mind interventions as well as drug prohibitions constitute interferences with freedom of thought which – if justifiable at all – can only be grounded in a paternalistic rationale. As one cannot remain uncommitted on these issues without engaging with limits of paternalism, only some aspects shall be briefly mentioned. Psychiatric interventions often restore mental capacities and thereby increase *internal* freedom of thought but at the expense of freedom from unwanted external interventions. Thus, one dimension of freedom of thought has to be balanced against the other. Insofar as treatments are restricted to (legally) incompetent patients under narrow conditions and in pursuit of their best interests (rather than those of society, relatives, or medical institutions), weighty ethical reasons seem to speak in their favor and therewith for an exception to the absolute status of freedom of thought.

Prohibitions of pharmaceuticals and other tools to change one's own mind beyond therapeutic contexts face similar challenges. Freedom of thought prima facie permits their use, but without doubt, any reasonable drug policy must install restrictive measures. However – and a bit provocatively – if the idea of freedom of thought was taken seriously, the prime aims of regulation would consist in *enabling* persons to safely alter their minds to enhance their internal freedom. Such an approach would seek to develop a regulatory system with the least restrictive constraints and apt harm-reduction strategies (medical supervision, substitution, drug checking, etc.). Prohibition (and criminalization) would be means of last resort to prevent severe self-harm. This perspective would constitute a paradigm shift. The unquestioned premise of today's drug policies lies in the eradication of consumption at almost all costs. In some countries, the harsh political climate against mind-altering substances begins to soften, partly because of the work of NGOs such as the high-profile "Global Commission on Drugs" which called for an end on the "war on drugs" in an influential report in 2011[20] as well as scientists who urge an evidence-based drug policy (Nutt et al. 2007; Nutt 2012). But as these are complex political decisions touching upon many social interests, it suffices to say here that the right to alter one's mind is not adequately

[20]Global Commission on Drug Policy 2011.

acknowledged in current regulatory schemes.A better test case for paternalism regarding alterations of one's own mind is provided by a recent case report of a DBS treatment of a patient suffering from anxiety and obsessive-compulsive disorder. The electrodes were placed in the nucleus accumbens, a part of the "pleasure center" or "reward system" of the brain (Cf. Kringelbach and Berridge 2009). Stimulation of this area elevates mood, in correlation with the increase in voltage: The higher the voltage, the stronger the mood elevation, up to a point at which the patient reported to be overwhelmed by euphoric, drug-like sensations of happiness. Stimulation also modulates, e.g., motivation and feeling of relaxation. A level of stimulation was found which put the patient in a normal state. Usually, patients can alter the intensity of stimulation through a control pad, but when the pleasure-center is targeted, different devices are installed to prevent overstimulation. As only the physician has access to the parameters of the stimulation, the patient asked him to turn the voltage up because he would like to feel "a bit happier" during the next weeks. The request was denied because of the limited knowledge over long-term effects. Uncertain himself, the physician put the perplexing questions where "therapeutic levels of happiness" run and whether stimulation should be confined to those levels to the medical community.[21]

The case raises intriguing questions for the law. By regulating well-being, feelings of pleasure, relaxation, and motivational states, the simulation targets central characteristics of the person. Presumably, this is the most invasive, precise, and direct form of accessing important psychological properties of another person currently available. Who should have the power to control these aspects? Freedom of thought is implicated because elevation of mood and motivation also alter, as case reports show, the kind of thoughts that come to mind. Other rights could be involved. If the stimulation is controlled through an external device (or a remote control), increasing or decreasing pleasure or altering mood against the will of the person likely violates human dignity. The present case is less infringing because the patient has consented to stimulation and merely desires its modification. The physicians contend they "are clearly not obligated to change DBS parameter settings beyond established therapeutic levels just because the patient requested it." They are right insofar as they do not have to actively participate in the patient's quest to alter his mind. But what if he only demands surrender of the controlling device of (or, depending on the technical realization, the means of access to) his own brain? The level of happiness is so central to the person that any denial of self-determination requires tremendously important reasons. After all, one should recall, the historically first inalienable right was the one to life, liberty, and the pursuit of *happiness*.

The case is structurally interesting as it throws the problem of legal paternalism over minds into full relief. Already spatially, the transformation occurs exclusively within the brain of the person and all effects likewise transform the inward domain

[21]See the full case report in Synofzik et al. 2012.

of consciousness only. Society or others are affected, just as Mill argued, only indirectly through the person herself and thus lack claims over the level of happiness.[22] Curbing access must thus be based on paternalistic grounds. In face of unforeseeable consequences of unrestricted self-stimulation of the reward system, the physicians' cautious approach is commendable. This leads to the paradoxical point that even though the matter is highly intimate and normatively only self-regarding, it is equally dangerous and in need of regulation. This case strongly suggests that even hard paternalism in regard to one's own fundamental interests is warranted.

All three examples of paternalism hint at a deeper challenge for neuroethics. What the law requires – and often uncritically presupposes – is an ethical assessment of the desirability of particular mental states, properties, and capacities. Paternalistic restrictions can be justified only if they promote long-term interests of the person, all things considered. Such assessments have to rely on judgments over the (dis-)value of particular mental states such as artificially induced blissfulness, symptoms of mental disorders, or effects of specific drugs. Otherwise, risks and benefits cannot be reasonably compared. In the DBS case, physicians tried to predict long-term consequences of heightened levels of happiness, e.g., its effects on wants and motivation in an attempt to define a therapeutic (why not optimal?) level of happiness. Similarly, mind-altering substances are often conceived as producing "high" states, but what if one rather speaks about enabling associative or creative thinking? Or, suppose the contention that some psychoactive substances as LSD facilitate insights into one's personality and exploration of otherwise unattainable psychological depths proves correct – which dangers are worth taking for it? A framework to address such points is currently missing. Many ethical elaborations of various mind interventions are ultimately based on moral considerations rather than long-term interests of individuals. The law evaluates mental states without a sound basis and is in need of assistance from an ethics of conscious (and other) mental states.[23]

Limits Public Interests

Presumably, the most controversial issue is neurointerventions without consent for public interests. Freedom of thought, in its current interpretation, opposes any such intervention. In spite of this, every state resorts to them. For instance, psychiatric patients are also treated to avert harm to others, not only to themselves. Some countries deploy psychoactive drugs for interrogation ("truth serum") (Calkins 2010) or to render defendants competent to stand trial. The – at least to European

[22]This becomes self-evident if one imagines a person with identical characteristics bestowed upon her on natural ways. Then, others surely do not posses any claims over them. Prima facie the same must be true for electrically modulated mood.

[23]Cf. the call for an ethics of consciousness by Metzinger 2009, Chap. 9.

ears – most cynical form of involuntary treatment is restoring convicts' "competence to be executed." These interventions do not paternalistically advance interests of affected persons, but of the state or society at large.

Further applications of neurointerventions for the public good are proposed, e.g., neurosurgery of criminals or moral enhancement of potentially every citizen, enhancing mental capacities of eyewitnesses or impairing their ability to fabricate lies, and, most worrisome, military uses of neuroscience (Moreno 2012; Marks 2010).[24] To clarify, as long as individuals consent to treatments (under appropriate conditions),[25] their rights are not infringed upon; only involuntary interventions are in dispute. Of course, courts have recognized an interest against being subjected to such interventions in cases that came before them, but they often held public interests to prevail. Broadly speaking, governments and the public have inconceivably many interests in the minds of citizens, but by themselves, interests in curbing freedoms do not provide a justification for doing so. The absolute nature of the right opposes balancing freedom of thought with public interests. The question is whether this strict prohibition should be upheld. To put it in a slightly different perspective: If one were to follow the temper of the times and allow certain interventions for the public good, one would have to establish mental duties of affected persons. Everyone who has to undergo mind interventions by law must be under a respective enforceable duty. Such duties contravene the idea that others do not have claims over the content of one's mind. Where those duties originate from and where they end would have to be explicated, which leads into fundamental issues over the general kind of duties persons owe to each other and to the state. To those endorsing consequentialist premises, strict limits to governmental powers irrespective of consequences appear utterly mistaken. Nonetheless, especially human rights law is founded upon the conviction that certain parts of persons are inviolable and have to remain outside of governmental reach.

While this matter cannot be settled here and deserves profound inquiry, it should be reminded that imposing mental duties constitutes a paradigm shift with potentially far-ranging implications. Instead of taking persons as they are and restricting their external freedom, the state acquires claims over the person herself. Instead of governing conduct, states govern minds. Is this not opening Pandora's box? Is it not, in times of omnipresent security worries, only a small step to policing thoughts, enforcing mental obedience, drugging the dissident? Even if one were to accept an expansion of governmental powers into the mental realm, one principle of public law will remain: States have to choose the least restrictive measure in curbing liberties. Oftentimes, curbing external freedom is less infringing than intervening into minds as freedom of action weighs much less than freedom of thought.

[24]Cf. the chapter on moral enhancement in this volume.

[25]Cf. the paper by Shaw in this volume.

Finally, a different set of public concerns can be found in ethical debates. It might be summed up as the wish of many people worried about the pharmacologization of everyday life (or, as David Healy put it: birth, Ritalin, Prozac, Viagra, death) to live in a natural society that does not medicalize its problems. As such, even if embraced by a vast majority, this (understandable) wish over a societal atmosphere cannot overrule freedom of thought. However, freedom of thought might be invoked at certain instances to counteract some societal developments.

The Right to Use Versus The Right to Refuse

Mill and Kant argued that mental alterations fall outside the purview of legal regulation because apart from deeds, they do not harm others or do not curb other persons' freedoms, respectively. The debate over neuroenhancements demonstrates that these assumptions might be misleading (CF. Greely 2006; Merkel 2007). In a mental economy, economic surplus is generated mainly through cognitive labor. In a sense, this realizes an old Marxist ideal: The means of production belong to the workers. Yet as long as they compete in job markets, this does not lead to liberation, but produces new forms of pressure. Optimizing the brain through neuroenhancements is optimizing the means of production. Legally, freedom of thought implies the right to use enhancements but by the same token to refuse their use. Therefore, opponents such as transhumanists and bioconservatives can – and should – both embrace the right, only emphasizing different dimensions.[26] Although strictly speaking, no one is coerced to take enhancements simply because others do, factual pressures may become so strong that persons cannot refuse without taking severe negative burdens upon them. This social conflict primarily results from constitutions of minds, not actions, and the different interests have to be reconciled in some way by a democratic legislator. Thus, even though it might appear paradoxical at first glance, one of the strongest reasons to curb freedom of thought of potential users is freedom of thought of refusers.

Future Directions

The seductive idea that thought is free can no longer serve as a basis for the way the law deals with minds. Neurotechnologies that afford to modulate and manipulate cognitive and emotional processes raise manifold challenges for the law: The right to freedom of thought has to be acknowledged and interpreted in a way that lives up to its theoretical and historical significance. Scope, contours, and interferences have to be defined. The current legal overemphasis on the side of

[26]For a detailed argument why bioconservatives should embrace cognitive liberty, see Bublitz 2013.

the sender of stimuli through, e.g., strong protection of freedom of expression has to find its limits in cognitive liberty as the right to remain free from unwanted manipulative interferences. This could place novel boundaries even on some socially accepted interventions. Perhaps most importantly, the absolute nature of the right has to be called into question. Very important public interests may justify interferences. If exceptions were allowed, the state would acquire the power to impose mental duties by which governmental competencies were expanded from controlling behavior to governing minds. This is a dramatic shift that requires profound debate. Furthermore, the law tacitly relies on extralegal reasoning about the desirability of mental states for which neuroethics could provide valuable assistance. Which mental states should states promote, discourage, or even outlaw? In the end, the scope of freedom of thought defines the realm and limits of mental self-determination or the boundaries of social access to the mind. Answers to these questions will have far-ranging repercussions in many areas of the law as well as in social and private life.

Cross-References

- ► A Duty to Remember, a Right to Forget? Memory Manipulations and the Law
- ► Ethics in Psychiatry
- ► Neuromarketing: What Is It and Is It a Threat to Privacy?
- ► The Morality of Moral Neuroenhancement
- ► The Use of Brain Interventions in Offender Rehabilitation Programs: Should It Be Mandatory, Voluntary, or Prohibited?
- ► Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Blitz, M. J. (2010). Freedom of thought for the extended mind: Cognitive enhancement and the constitution. *Wisconsin Law Review, 4*, 1049–1117.

Boire, R. G. (1999). On cognitive liberty I. *Journal of Cognitive Liberties, 1*, 7–13.

Boire, R. G. (2000). On cognitive liberty II. *Journal of Cognitive Liberties, 2*(2), 7–20.

Boire, R. G. (2001). On cognitive liberty III. *Journal of Cognitive Liberties, 2*(1), 7–22.

Boire, R. G. (2005). Neurocops: The politics of prohibition and the future of enforcing social policy from inside the body. *Journal of Law and Health, 19*(2), 216–257.

Bublitz, J. C. (2013). My mind is mine!? Cognitive liberty as a legal concept. In E. Hildt (Ed.), *Cognitive enhancement* (pp. 233–264). Dordrecht: Springer.

Bublitz, J. C., & Merkel, R. (2014). Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. *Criminal Law & Philosophy, 8*(1), 51–77.

Bury, J. B. (1952). *A history of freedom of thought*. New York: Oxford University Press.

Calkins, L. (2010). Detained and drugged: A brief overview of the use of pharmaceuticals for the interrogation of suspects, prisoners, patients and POWS in the US. *Bioethics, 24*(1), 27–34.

Delgado, J. (1969). *Physical control of the mind. Toward a Psychocivilized Society*. New York: Harper & Row.
Eide, A., & Swinehart, T. (1992). *The universal declaration of human rights. A commentary*. Oslo: Scandinavian University Press.
Global Commission on Drug Policy. (2011). Report “War on drugs”. www.globalcommissionondrugs.org
Greely, H. T. (2006). The social effects of advances in neuroscience: Legal problems, legal perspectives. In J. Illes (Ed.), *Neuroethics: Defining the issues in theory, practice and policy*. Oxford/New York: Oxford University Press.
Harris, D., O’Boyle, M., & Warbrick, C. (2009). *Law of the European convention on human rights* (2nd ed.). Oxford: Oxford University Press.
Hobbes, T. (1996). In R. Tuck (ed.), *Leviathan* [1651]. Cambridge: Cambridge University Press.
Husak, D. (1992). *Drugs and rights*. Cambridge: Cambridge University Press.
Kant, I. (1991). *Metaphysik der Sitten* (Metaphysics of Morals, 1791). Cambridge; New York: Cambridge University Press.
Kringelbach, M. L., & Berridge, K. C. (2009). *Pleasures of the brain*. New York: Oxford University Press.
Levy, N. (2007). *Neuroethics*. Cambridge: Cambridge University Press.
Lifton, R. J. (1989). *Thought reform and the psychology of totalism. A study of brainwashing in China*. Chapel Hill: University of North Carolina Press.
Lipsman, N., Woodside, D. B., Giacobbe, P., Hamani, C., Carter, J. C., Norwood, S. J., et al. (2013). Subcallosal cingulate deep brain stimulation for treatment-refractory anorexia nervosa: A phase 1 pilot trial. *The Lancet, 381*(9875), 1361–1370.
Macklin, R. (1981). *Man, mind, and morality. The ethics of behavioral control*. Englewood Cliffs, N.J: Prentice Hall.
Marks, J. (1992). *Search for the Manchurian candidate. CIA and mind control*. New York: W.W.Norton.
Marks, J. H. (2010). A neuroskeptic’s guide to neuroethics and national security. *AJOB Neuroscience, 1*(2), 4–12.
Maslen, H., Douglas, T., Kadosh, R. C., Levy, N., & Savulescu, J. (2014). The regulation of cognitive enhancement devices: extending the medical model. *Journal of Law and Biosciences, 1*(1), 68–93.
Merkel, R. (2007). *Intervening in the brain. Changing psyche and society*. Berlin: Springer.
Metzinger, T. (2009). *The ego tunnel. The science of the mind and the myth of the self*. New York: Basic Books.
Mill, J. S., Bromwich, D., Elshtain, J. B., & Kateb, G. (2003). *On liberty*. New Haven: Yale University Press.
Moreno, J. D. (2012). *Mind wars. Brain science and the military in the twenty-first century*. New York: Bellevue Literary Press.
Nowak, M. (2005). *U.N. Covenant on Civil and Political Rights. CCPR commentary* (2nd ed.). Arlington: N.P. Engel.
Nutt, D. J. (2012). *Drugs – without the hot air. Minimising the harms of legal and illegal drugs*. Cambridge, UK: UIT.
Nutt, D., King, L. A., Saulsbury, W., & Blakemore, C. (2007). Development of a rational scale to assess the harm of drugs of potential misuse. *The Lancet, 369*(9566), 1047–1053.
Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews. Neuroscience, 13*(8), 587–597.
Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science, 342*(6157), 482–484.
Sargant, W. W. (1997). *Battle for the mind. A physiology of conversion and brain-washing*. Cambridge, MA: Malor Books.
Schlaepfer, T. E., Cohen, M. X., Frick, C., Kosel, M., Brodesser, D., Axmacher, N., et al. (2007). Deep brain stimulation to reward circuitry alleviates anhedonia in refractory major depression. *Neuropsychopharmacology, 33*(2), 368–377.

Sententia, W. (2004). Neuroethical considerations: Cognitive liberty and converging technologies for improving human cognition. *Annals of the New York Academy of Sciences, 2004*, 221–228.

Spinoza, B, de. (2007). In J. Israel (ed.), *Theological-political treatise* [1670]. Cambridge: Cambridge University Press.

Synofzik, M., Schlaepfer, T. E., & Fins, J. J. (2012). How happy is too happy? Euphoria, neuroethics, and deep brain stimulation of the nucleus accumbens. *AJOB Neuroscience, 3*(1), 30–36.

Taylor, K. E. (2006). *Brainwashing. The science of thought control*. Oxford/New York: Oxford University Press.

Tussman, J. (1977). *Government and the mind*. New York: Oxford University Press.

Valenstein, E. S. (1974). *Brain control. A critical examination of brain stimulation and psychosurgery*. New York: Wiley.

van Dijk, P., & van Hoof, G. J. H. (2006). *Theory and practice of the European convention on human rights* (4th ed.). Antwerp: Intersentia.

Neuroimaging and Criminal Law

84

Reinhard Merkel

Contents

Abstract

Methods of neuroimaging have sporadically, though in recent years increasingly, occurred in legal proceedings. By now, however, it seems that they are about to enter courtrooms on a systematic basis. This poses a host of normative problems, to do, for instance, with future applications of neuroimaging to determine culpability, to test the veracity of testimony, or to predict the future dangerousness of perpetrators. The latter two, brain-based lie detection and “neuroprediction” of dangerousness, are examined in this chapter. Functional

R. Merkel
Faculty of Law, University of Hamburg, Hamburg, Germany
e-mail: reinhard.merkel@jura.uni-hamburg.de

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_165,

magnetic resonance imaging (fMRI) is taken as a paradigm model, and its potential impacts on criminal trials are explored. The analysis is premised on a range of basic distinctions: between (1) different phases of a criminal trial; (2) the divergent roles played by the parties to a trial, most notably prosecution and counsel, and the different evidentiary goals and burdens associated with these roles; and (3) between compulsory and consensual fMRI. It turns out that there are no good reasons to ban fMRI for lie detection or for neuroprediction from criminal proceedings entirely. Instead, it should be admitted differentially in criminal trials, viz., only for purposes of exoneration, but not of conviction, of the defendant. Substantiating arguments are expounded. In cases of preventive detention, it may even be obligatory for the state to offer chances of possibly exonerating brain imaging to perpetrators who were otherwise considered candidates for indefinite custody.

Introduction

Unless all indications are deceptive, we are at the brink of a more or less systematic introduction of various methods of neuroimaging into legal proceedings.[1] This pertains to all major areas of the law, public as well as civil and criminal law.[2] Here I will focus on criminal law and its normative foundations only. Indeed, it seems fair to say that, for a variety of reasons, none of the other legal spheres will be affected as profoundly and in a comparably vexing manner by the suggested development as the substantive and the procedural law of crimes. This needs no elaboration here. Suffice it to say that in criminal law the state deploys its sharpest sword, as it were, against its own citizens, and hence has a constitutional duty to meet the most stringent obligations to justify its measures. Broadly speaking, this is the main reason why any uncertainty concerning validity and reliability of methods of neuroimaging as well as its potential to intrude into protected spheres of its subjects is bound to cause greater irritations in the field of criminal law than in any other legal area.

Over the last decades, a whole range of brain imaging methods have been developed or, if already on hand, significantly enhanced. For the purposes of our analysis, their classification into structural and functional techniques is of particular interest. Those of the former type provide ways of demonstrating the anatomical status of the brain, whereas the latter make its activities (functioning) accessible to outside observers. Structural procedures are, for instance, CT (computed tomography) and MRI (magnet resonance imaging); functional methods are electroencephalography (EEG), positron emission tomography (PET), single-photon emission

[1]This, of course, is a very general and hence not very informative statement. Before it can be elaborated on, it needs to be differentiated into a host of distinctions pertaining to conceptual as well as factual and normative questions – matters which I will deal with in due course.

[2]Least so, presumably, to public law, for reasons not to be developed here. On the various potential applications in civil law (see Granacher 2012; Eggen and Laury 2012; Moriarty et al. 2013).

computed tomography (SPECT), and functional magnetic resonance imaging (fMRI). In what follows, I will concentrate on methods of fMRI. They are currently the most technically advanced and, in various respects, the most promising ways to explore brain functions. Hence, the normative problems of neuroimaging that we will focus on can best be expounded by examining this paradigm model.

I will not talk about the clinical merits of brain imaging, i.e., its diagnostic, therapeutic, and preventive values which are largely beyond dispute. Instead, I will investigate some of its nonclinical applications. All of them have long become objectives of extended research in basic as well as in applied sciences. But some of them have also entered the courtroom. That is what makes them normatively interesting in various respects, some of which appear to be new and unique.

Basic Science: Techniques

As indicated, magnetic resonance imaging (MRI) in its standard forms can be either structural or functional, depending on what it is destined to measure: anatomical structures of a brain or certain of its activities, viz., functions (Bigler et al. 2012). Structural images have their primary field of application in clinical contexts of diagnostic or therapeutic provenance.[3] More interesting for the purposes of criminal law as well as legal ethics is functional magnetic resonance imaging (fMRI). For according to the expectations commonly associated with its use, it is supposed to make observable to external examiners, though only in an indirect way, the neural correlates of certain mental processes in the minds of persons tested in a scanner while being engaged in various cognitive tasks or exposed to stimuli designed to evoke emotive reactions. And this is said to procure substantive ground from which the goings-on in subjects' minds become accessible (deducible) in some sense for outside observers. Sure enough, most of this wording amounts to claims that are much too sweeping. They require several conceptual and technical qualifications as well as normative *caveats*. We will turn to these later.

The physics and technical intricacies of fMRI are very complex, much more so than could be surmised by looking at what the media tell us about it, with their basic microphysical processes rooted in fundamentals of quantum mechanics. For our purposes here, we need not delve into such difficulties. The following sketchy account may suffice.[4] FMRI is an imaging technique that records certain "changes

[3]They do, of course, also have various applications in the law, e.g., in civil cases, such as medical malpractice suits where they might be used to demonstrate differences in structural brain conditions before and after an intervention, etc. Normatively, these are mostly trivial problems of no particular pertinence to imaging methods.

[4]For a brief yet still formidable description see Logothetis 2008; extensively Logothetis and Pfeuffer 2004; for overviews accessible to laypeople see Raichle 2009; Jones et al 2009; Langleben et al. 2012; Taylor 2012, pp. 103–110.

in the regional blood volume and flow that are associated with cognitive activity" (Langleben and Moriarty 2013). Oxygen consumption, glucose utilization, and therefore blood flow in the brain constantly change in ways that are closely and in a remarkably precise mode related to cellular activities in the respective neural areas (Raichle 2009, p. 118). Such cellular activities, in turn, are believed to be the substrate, and hence (in some sense) the physiological mirror, of mental processes of all kinds in the minds of subjects during their test in a scanner.

The data acquired in fMRI are based on differences in the magnetic properties of the content of blood vessels and surrounding tissues in the brain as well as on such differences between oxygenated (aortic) and deoxygenated (venous) blood. These differences are elicited and made measurable by exploiting differences in the "resonant" properties of subatomic particles (protons) of hydrogen atoms contained in the water molecules of brain cells and surrounding blood flows, namely, their varying capacities to resonate (i.e., to absorb and emit energy quanta) while being exposed to brief pulses of radiation. This is achieved by placing the head of the subject into a powerful magnetic field in the range of 0.5–5 T, generated and modified by the scanner, thus uniformly aligning all magnetic particles in the targeted brain areas, and then emitting large numbers of very brief consecutive radio waves which are also produced by the scanner. These pulses deflect the previously aligned magnetic axes of the hydrogen protons away from their state of collective equilibrium. Soon as a radio pulse stops, the protons return to their former equilibrium, thereby "resonating," i.e., emitting miniscule radio quanta that can be detected by the scanner. These signals vary in strength depending on which type of tissue the emitting protons belong to. Thus, they characterize the diverse organic matter from which they originate: oxygenated or deoxygenated blood or other tissue. It is these differential radiation echoes that the scanner records and processes in order to create its fMR images (cf. Matthews and Jezzard 2004).

The specific brain activity believed to be the substrate or (at least) correlate of certain mental activities is indicated chiefly by the varying "echoes" of either oxygenated or deoxygenated blood particles, hence the label these signals are commonly marked with: "blood oxygenation-level dependent" (BOLD). They are of particular interest to researchers or clinicians. BOLD imaging is presently the most widely used fMRI technique (Langleben and Moriarty 2013, p. 223).

The scanner also permits localizations of the collected signals in (brain) space, endowing the method with a high quality of spatial resolution.[5] By applying large quantities of radio pulses from different angles all around the head and recording

[5]Temporal resolution of fMRI is of necessity weaker since the physiological reactions of the brain to the various tasks presented to its bearer, i.e., the increased blood flow to its respective parts, take a few seconds time.

the diverse echo pulses thus elicited, huge amounts of data are collected by such successive measurement "slices" through the brain. The scanner further subdivides these slices into relevant units of volume, so-called voxels,[6] tiny cubicles of (currently) 1–3 mm per side. Data derived from the voxels can then be configurated by a computer program into 3D images of the brain. Subsequently, these images are colored differentially by the computer and projected onto a baseline picture of the individual brain, usually taken beforehand by a structural MRI. The colors used in this process accord to the varying quantities of blood flow indicated by the recorded BOLD signals. They have no inherent meaning and do not mirror a colorful reality in the brain. It is the researchers that (to some extent arbitrarily) specify what, i.e., which amount of differential neural activity, is expressed by each color. By convention, the general rule is that "the brighter the color (say yellow compared to orange) the greater the statistical significance of the differences between two conditions" (Jones et al. 2009).

It is important to note that these data do not depict any "absolute" measure of regional brain activity. Rather, what they indicate is "relative regional activity over time" during the test (Langleben and Moriarty 2013, p. 223). This is achieved by relying on a method called "cognitive subtraction" (Aguirre and D'Esposito 1999). Here is a brief expert description:

> This principle assumes that the fMRI signal difference between two behavioral conditions that are identical in all but a single variable is due to this variable. Therefore, a proper comparison (i.e., control) condition is critical for meaningful BOLD fMRI paradigm. The fMRI activation maps reported in the literature usually represent a statistical subtraction between the fMRI activity maps related to the target and control variables. Ideally, the comparison and target conditions would be identical, except for a single variable of interest (Langleben and Moriarty 2013, p. 223; see also Raichle 2009, p. 5).

Reliance on the "cognitive subtraction" method demands observance of elaborate criteria for highly sophisticated experimental designs on the part of the researchers if their endeavor is to be successful in terms of clinical relevance or meaningful in terms of basic science.

Metaphysical Problems of Mind and Brain?

This method, however, also accounts for quite another fundamental aspect of fMRI. Philosophers who have read their way through the vast literature on fMRI might wonder why there is never any mention, not even in passing, of the age-old metaphysical problem of the relationship between the mind and the brain. If one aspires to measure (albeit indirectly) the brain activity of a person in order to disclose certain mental processes in their minds, does that not imply that one adheres to a certain type of monistic materialism (or physicalism) in the said metaphysical debate, that is, to a position holding that all mental

[6] Derived from "volume element" in assonance with the two-dimensional visual unit "pixel."

events are either simply identical to ("nothing over and above") correlated processes in the brain or yet "supervene" on such processes and hence metaphysically depend on them? And is that simply to be taken for granted? In other words, aren't there any reputable dualisms around anymore in contemporary metaphysics?[7]

We may, however, leave this issue aside here as unresolved as it is in philosophy. The "cognitive subtraction" principle of fMRI makes sure that, irrespective of fundamental mind-brain metaphysics, fMRI scans are apt to deliver empirical results that can reasonably be taken as disclosures of relevant facts about the mind. For what they aim to determine is not some "absolute" mental state "in itself," plainly read off its neural substrate. That would certainly amount to a more magical type of science fiction. Rather, they only seek to establish a difference between two brain states by comparing two different patterns of brain activity. One of these is the "baseline" (or "control") state before the test person is confronted with a particular cognitive task, whereas the other is task related, i.e., induced by the particular variable in the experimental setting which is introduced by the test-specific performance. If a difference between the brain activation patterns of the baseline state and the task-related state consistently emerges while the subject performs a specific cognitive task, one can be reasonably sure that the variation in those mental states manifests itself somehow in the recorded difference in brain patterns and can thus be identified by the latter. Put another way, that the resulting "difference image" would represent only those neural areas concerned with the specific task of the particular test (Raichle 2009, p. 5). This is quite independent from whatever else the relation between mind and brain might be in terms of metaphysical monism, reductionism, dualism, supervenience, or even Cartesian causal interactionism. Hence, we can safely lay aside these unresolved problems of the philosophy of mind.

Neuroimaging of Deception: Feasibility? Admissibility?

As indicated above, our normative analysis has to start with some fundamental distinctions.

1. Firstly, one should keep apart the different purposes for which fMRI could possibly be utilized in criminal law procedures. Three of them are rather obvious

[7]For elaborated overviews on mind-brain identity theories, on supervenience and on dualism(s), see Stanford Encyclopedia of Philosophy, entries "Mind/Brain Identity Theory" (Smart 2013), "Supervenience" (McLaughlin and Bennett 2013), and "Dualism" (Robinson 2013). To be exact, supervenience theories, holding that mental processes are asymmetrically dependent on corresponding brain processes, are really a form of moderate dualism (albeit decisively different from all types of "substance dualism" in the wake of Descartes' classical position which has only few followers these days). Supervenience, of course, is perfectly compatible with the assumption that measuring certain brain processes via fMRI can disclose correlated mental events.

(albeit not exhaustive): lie detection, assessment of responsibility, and prediction of future dangerousness. In what follows, I will confine my analysis to the first and last of these issues, i.e., on neurodetection (of lies) and neuroprediction (of dangers).[8]

2. Secondly, the question whether fMRI is a sufficiently reliable method of evidence in criminal law and hence admissible in the respective legal proceedings should be distinguished from the normative question whether such applications can be justified in principle, which is to say, even if fMRI turned out to be a scientifically valid and reliable method of evidence. It is by no means clear ex ante which of the risks possibly posed by the use of neuroimaging in criminal law appears more alarming: that of their uselessness or of their potency for evidentiary purposes.
3. Thirdly, the different roles of the parties in criminal trials[9] associated with divergent, in part conflicting, interests must be taken into account. Not all of these roles are equally suited for an attempt to deploy neuroimaging methods for one's particular aims.
4. Furthermore, a criminal trial is not a homogenous process with one invariable objective (be it retribution, prevention, a mixture of both, or something else) remaining constant through all parts of the proceedings, and with immutable interests of the parties involved as well as of the state and the public. Rather, it consists of a number of distinct phases, each of which is assigned a different principal purpose by the law and in each of which the roles of the parties and of the witnessing public also change to varying degrees.
5. Finally, of particular importance here is the distinction between the trial procedures in court (criminal proceedings proper, as it were) and, as the case may be, the subsequent enforcement of a prison term. During the latter too, insights into the mental sphere of a prisoner, possibly procured by an fMRI, might well be of special importance, be it to the prisoner himself or to the state and the public, for instance, if such insights were apt to sustain a reasonable prediction about the future dangerousness of the prisoner after a potential discharge from jail.

All of the above distinctions are of a more or less rough-and-ready kind. But they must be taken into account as one reflects on the problems of fMRIs in criminal procedures. All of them are associated with varying kinds and degrees of significance that the results of a neurotechnical access to the mental sphere of a party, be it the defendant or a witness, might have for that party itself as well as for the others in

[8]The important issue of neuro-assessing guilt and responsibility would require a whole treatise of its own, presupposing a clarification of the concept of (criminal) responsibility, which in turn implies hotly contested issues of freedom of the will, including the "principle of alternative possibilities" (PAP) as an alleged prerequisite of free will and responsibility, and other perennial issues in the metaphysics of mind. I have developed my own thoughts on these topics elsewhere (cf. Merkel 2008, 2011).

[9]Mainly the accused and his or her counsel; the prosecutor, judges, and jury; witnesses and expert witnesses; and, to a limited extent, the victim of the tried criminal offence.

every phase of the process. Put in the sweeping form of whether or not neuroimaging is admissible to courts in criminal law, the question admits no sensible answer. So the following considerations will keep in line with the foregoing delineations.

Lie Detection in the Trial Phase: Validity? Reliability?

For quite some time, measurements of physiological indicators of deception, such as skin conductivity, blood pressure, respiration, and a few more, have been conducted by employing the traditional "polygraph" which measures activity in the peripheral nervous system to detect deception. Two different methods of acquiring the relevant data have been used: the "control question test" (CQT) and the "concealed information test" (CIT), or more popular "guilty knowledge test" (cf. MacLaren 2001). In the course of the former, usually three types of questions are posed to the test person: firstly, incriminatory ones (such as "Did you kill your wife?"); secondly, irrelevant and harmless control questions (such as "Who is the current president of the United States?"); and thirdly, control questions about nonspecific misconduct, designed to somehow strain even nondeceptive subjects, but not in an specifically inculpatory way, i.e., not with regard to the particular incident in question (such as "Have you ever mistreated your wife or your children?"). By contrast, the CIT takes aim at highly specific and crime-related knowledge that subjects could only possibly have if they were indeed involved in the perpetration of the criminal offence in question. If that is the case, their autonomic nervous system will respond differently from that of an innocent person to such interrogating stimuli. Or so it is claimed.

The polygraph has come into disrepute, at least with regard to its application in criminal cases. Most, though not all, courts in Europe as well as in the United States have disapproved polygraph-based evidence (cf. National Research Council 2003, Chaps. 3 and 4). This pertains particularly to its CQT version as opposed to the CIT. The latter method whose reliability appears to be significantly superior to that of CQT (Ben-Shakar and Elaad 2003, p. 132; MacLaren 2001) has also been extensively examined in correlation with fMRI (Hakun et al. 2008; Gamer et al. 2007, 2012). Insofar, some of the prospects for a scientifically valid forensic application expressed by competent researchers have a considerably more optimistic tone than the majority of voices dismissing the polygraph (Gamer et al. 2012, p. 513; Langleben et al. 2012; Langleben and Moriarty 2013; see also Vincent 2011). This is, however, vehemently contested terrain. Other scholars, by contrast, are no less concerned about fMRI applications in court than they are about polygraphy (Mobbs et al 2007). Some of them resolutely deny an admissibility of fMRI lie detection in court at least for the time being (Morse 2006, 2012; Greely and Illes 2007; Sinnott-Armstrong et al. 2008; Uttal 2009; Pustilnik 2009; Kanwisher 2009; Moriarty 2009; Brown and Murphy 2010; Oullier 2011; further references in Schauer 2010, pp. 1199–1202). Others even demand an outright regulatory ban on "any non-research use of new methods of lie detection, including specifically

fMRI-based lie detection, unless or until the method has been proven safe and effective to the satisfaction of a regulatory agency and has been vetted through the peer-reviewed scientific literature" (Greely and Illes op.cit. 413). Furthermore, to some extent, the controversy is reflected in a whole range of court decisions. Some of these have admitted fMRI in criminal proceedings (though mostly to no avail for the verdict), whereas (more) others have rejected it.[10] It appears impossible, at least for legal scholars, to take a fair, scientifically attestable and yet decisive stand on either one of the sides in this controversy.

There is, however, also no need to do so here in order to sensibly pursue the matter further. Most of the concerns voiced by skeptics relate to certain standards of scientific validity and reliability in expert testimony required for evidentiary purposes.[11] With regard to such standards, they find all currently available methods of fMRI for lie detection to be wanting in various respects. And from this perspective, a plausible prima facie case against fMRI lie detection can be made. Objections may invoke not only the technical difficulties of the scanning procedure or the intricate problems of developing effective test paradigms but also the complexity of the cognitive processes involved in the natural phenomenon to be investigated: human deceptive behavior.

Such behavior includes cognitive processes to generate intent and strategies of deception in a given context as well as executive processes to perform the chosen deceptive act (Luber et al. 2009). A useful, albeit still coarse, taxonomy distinguishes four categories of cognitive functions associated with deception: information management, risk management, impression management, and reputation management (Sip et al. 2007). Each of these activities itself encompasses a set of more elementary functions, such as planning, employing one's working memory, selecting between alternatives, and modulating response inhibition (Luber et al. 2007). All of these processes are assumed to interact in a systematic, though largely unconscious, way during deceptive behavior. Furthermore, in order to be accessible to fMRI investigations, this complex cognitive machinery must, at least roughly, be understood in terms of its neural underpinnings – a formidable task fraught with nescience and uncertainties. To cut this still extendable account somewhat short, if one simply assesses the potential of fMRI to disclose deception vis-à-vis such difficulties and against the backdrop of a universal standard of validity and reliability in natural sciences, the prospects for success certainly are rather bleak.

[10]Overviews in Jones and Shen 2012; Aronson 2010. Most noteworthy of these rulings perhaps the verdict in *Semrau* (US v. Semrau 2010; affirmed US Court of Appeals, 6th Cir. 2012; see also Gary Smith v. State of Maryland 2012). It is not only based on established juridical reasoning but also on an in-depth examination of the available scientific evidence from the perspective of evidentiary requirements in the criminal law.

[11]In science, "validity" and "reliability" are distinguished (though related); the former refers to whether research results really demonstrate what they purport to, whereas the latter to whether such results are consistently obtained in sufficiently equal experimental situations. In law, however, both terms are frequently used interchangeably. Nothing hinges on this terminology here; so in what follows, "validity" can usually also be read as "reliability," and vice versa.

There is, however, no such thing as a uniform, homogenous standard of scientific validity applying to all evidentiary purposes that are legitimately pursued by parties to criminal proceedings. This pertains especially to the three major players in court: the defendant, the prosecution, and the jurors or the judges. As is well known, the defendant, or his counsel, does not have to proof anything, neither his innocence nor truthfulness. Things are quite different from the perspective of prosecutors and, in the end, judges or juries. What stands in need of sound proof in criminal trials is every single precondition required for a guilty verdict: firstly, all relevant facts establishing that the deed in question was committed by the accused[12]; secondly, the absence of justifying circumstances, such as self-defense[13]; and finally, the preconditions of the defendant's personal culpability.[14] All of this must be established "beyond reasonable doubt" or (amounting to the same standard) "to the firm conviction of the court" by prosecutorial evidence. And none of its potential refutations falls on the part of the accused, not even to some weak standard of plausibility, let alone doubtlessness. On the other hand, he or she certainly has a right to proffer all sorts of possibly refuting evidence that might exonerate them. This decidedly asymmetric allocation of the burden of proof, a corollary of the fundamental principle of "*in dubio pro reo*" ("benefit of the doubt" for the suspect),[15] must have, one would surmise, some bearing on the question of standards of scientific validity that have to be met by the prosecution, on the one hand, and by the defendant, on the other, in their respective efforts to provide evidence for or against certain relevant circumstances.

Against this background, all a defendant needs to establish in order to evade conviction is some qualms in jurors' or judges' "firm" confidence in his guilt to the rather small degree that suffices to preclude the "beyond doubt" criterion. So he

[12]Which encompasses all objective facts constituting the commission of the offence ("actus reus") as well as the required subjective facts on the part of the defendant ("mens rea").

[13]Note that this absence of justifying circumstances is regularly *presumed* unless there are concrete indications to the contrary. In other words, the realization of the elements of crime of a particular offence, i.e., of a behavior that is *generally* forbidden on pain of punishment, indicates prima facie that it was also forbidden (unjustified) in the *concrete* case. So with regard to particular justifications, there is often no need to proof anything. But if there are indicia of the presence of justifying circumstances during the perpetration of the deed, the final burden of proof of their *absence* falls on the prosecutor.

[14]And, as the case may be, the absence of specific exculpatory circumstances. Not all legal orders contain provisions for such specific exculpations, but many do. For instance, according to Art. 17 of the German Criminal Code (GCC), an *inevitable* error on the defendant's part that his behavior is (or was) lawful exculpates him (but does not, of course, make the *deed* itself lawful); certain forms of non-justifying necessity (threats to life, bodily integrity, and personal freedom) also have this effect (Art. 35 GCC).

[15]Contained in various Human Rights Conventions in international law, e.g., Art. 6 para 2 of the European Convention of Human Rights (1950).

may well be interested in submitting evidence of his truthfulness whose scientific reliability appears rather shaky – if and when it is at least above a threshold of mere "junk science."[16] There is no doubt that at present the results of research in fMRI lie detection are already way above this threshold (even if, as yet, they have been derived almost entirely from basic research in laboratories and scarcely been applied to real-life situations). Thus, it can be fully appropriate for a defendant to proffer methods of expert testimony in favor of his veracity that the prosecution could not sincerely think of to proof the opposite. Such measures may well fit his own aims, while being entirely inept and hence inadmissible for prosecutorial purposes of determining guilt. The weak or contested scientific status of such methods of evidence would then be simply a matter of assessing their correspondingly weak probative value, but not a matter of their (in)admissibility.[17]

As to the polygraph, however, criminal courts in the United States as well as in Europe have tended to decide differently.[18] And with regard to the admissibility of fMRI, courts at least in the United States seem to more or less consistently stick to the principles of their reluctant judicature on polygraphy. That is to say, they appear to apply rather uniform standards of scientific reliability to expert testimony regardless of whether that evidence is obtained and assessed for purposes of convicting or of exonerating the accused. Hence, courts seem to ignore the fundamental difference in probative aims and burdens associated with prosecutorial purposes (and duties), on the one hand, and defense goals, on the other.[19]

Dangerous "Seductive Allures"?

Why this is so is not entirely clear. In its landmark decision *Daubert v. Merrell Dow Pharms, Inc.* (1993), the US Supreme Court recognizes that "scientific validity for one purpose is not necessarily scientific validity for other, unrelated purposes" (1993, at 591). This pertains to possible varieties of diverse *scientific* topics whose proof might be the aim of expert testimonies. But why not also apply this

[16]A criterion derived, e.g., from the US Federal Rule of Evidence 702; see *Best v. Lowe's Home Centers, Inc.*, (2009), at 176; a comparable threshold in Art. 244 para. 3 of the German Criminal Procedures Act ("entirely unfit and useless"). However, in accordance with the asymmetry principle in the evidence pointed out in the previous paragraph, the threshold of what counts as "junk science" also varies with regard to what evidentiary goal is at stake: either conviction (prosecutorial) or exonerating (defense-related) purposes of the evidence at hand.

[17]Of course, expert testimony introduced for either prosecutorial or defense purposes may ex post turn out to back up just the opposite purpose, respectively. Then, it must be taken by its probative value for what it objectively supports. Evidently, however, the question of admissibility must be decided ex ante: by assessing the probative value of the evidence for the aim it is intended to further by the party who proffers it.

[18]For the United States see Schauer 2010, p. 1196, n. 23; for Germany and, in passing, a few other European countries see Putzke et al. 2009.

[19]As to the jurisdiction in the United States, cf. the critical voices in the literature cited above; as to Germany, see, for instance, BGHSt 44, 308 (1998).

uncontested rule to the diverse *evidentiary* aims that the parties of a trial pursue? In both these situations, what is at stake are divergent goals which are supposed to be supported by the evidence proffered respectively. If it is true that "the same evidence may be extremely probative for one purpose and not even relevant for another,"[20] why not also concede that it may well be of a (albeit small) probative value for the purposes of one party and not even relevant for those of another?[21] Granted that the state of the art in fMRI for lie detection is "not anywhere near meeting the *Daubert* standard" (Seaman 2009, p. 933). It does not follow, however, that its use in court, if exacted by the appropriate party for their particular goals, is excluded. This not only accords with our considered judgments but also is corroborated by our clear intuitions. Consider the infamous decision of an Indian court that sentenced a woman to life imprisonment for murder in 2008. The verdict was crucially based on evidence from BEOS (brain electrical oscillations signature), a form of EEG brain scanning. It rightly strikes most researchers as well as most legal scholars as outrageous (cf. Giridharadas 2008; Aggarwal 2009). Now, by contrast, consider an acquittal of a suspect based on neuroimaging, for instance, the (hypothetical) ex post acquittal of the woman who, in a recent UK case, was convicted of poisoning a child in her care.[22] She served a prison term of several years but continued to assert her innocence. After her discharge, she was examined in a series of fMRI scans while being confronted with her own and, on the other hand, with the public prosecutor's version of the indicted incident. The results were markedly consistent with her own story. Thus, the scans, while certainly not "proving" that she was innocent, notably demonstrated that she may have been (cf. Spence et al. 2008). Few would find it scandalous had the case been retried on this new evidentiary basis and had the defendant then been acquitted on the "benefit of the doubt" principle.

Still, the postulate of differentiating the applicable standards of validity with regard to who demands an fMRI for what purpose in the trial does not meet with much approval by courts and many scholars. The reason for this seems to be something that is captured, for instance, in the US Federal Rule of Evidence 403. It permits courts "to exclude even relevant evidence if its probative value is substantially outweighed by a danger of confusing the issues or misleading the jury" (cf. US v. Semrau 2012, p. 17). Such a confusing and misleading potential is seen in the frequently invoked "seductive allure" that neuroimaging is believed to unfold vis-à-vis laypersons such as jurors in criminal trials (Weisberg et al. 2008). It has been termed "Christmas tree phenomenon" (Mobbs et al. 2007), alluding to the glittering colors of the computer-generated fMRI pictures indicating the differential blood flow in the examined brain areas. The idea seems to be that jurors or even judges might somehow be overwhelmed by such suggestive effects in their

[20]Brown and Murphy (2010, p. 1155), endorsing the cited evidentiary rule in *Daubert*.

[21]Cf. Merkel (2011, 244 pp.), see also the excellent exposition in Schauer (2010).

[22]The case was not retried, so there has not been an actual ex post acquittal of the convict, but it is not at all far-fetched that there could have been one.

ability to soberly assess fMRI evidence; hence, the admission of fMRI evidence would fly in the face of the Federal Rule of Evidence 403 (Sinnott-Armstrong et al. 2008).

Such concerns have an initial plausibility.[23] On second thoughts, however, they appear rather puzzling. If fMRI can provide relevant evidence for defense purposes, its admissibility is prima facie backed by a fundamental, in some jurisdictions constitutional, right of the defendant. This is obviously a matter of profound legal importance. If, on the other hand, there is a risk of seductive "Christmas tree" phenomena beclouding jurors' and even judges' capacities to soberly assess what's presented to them, why not then instruct jurors and judges how to avoid that risk and interpret the images properly (Feigenson 2006)? That is, why not explain the usually very limited probative value of such evidence even for the modest purposes of the defendant? Why preclude the entire option and thus run the risk of curtailing a defendant's basic rights?

One may suspect that another, rather clandestine intuition is at work here: one of fairness, stipulating a principle of "equality of weapons" between parties to a criminal trial. If the prosecution, so the idea might go, has no chance to introduce fMRI for purposes of conviction, then defendants should not be allowed either to avail themselves of such an opportunity for the opposite aim. But this again seems to rest on a misunderstanding of the roles of the parties in a criminal trial and the asymmetric tasks associated with these roles. From the perspective of the state and thus the public prosecutor, a criminal trial has (or, in any case, should have) nothing to do with a kind of adversarial contest which the prosecution should be determined to "win" by getting the accused sentenced and which therefore should be played on equal terms and premises.[24] Rather, a criminal trial is a procedure to publicly defend the validity of a broken social norm by making the person who is responsible for the breach "pay" for what they did, in order to symbolically reinstall ("repair") the norm's violated normative claim to universal obedience. Against this backdrop, the above-sketched asymmetry in evidentiary requirements between prosecutors and defendants gains its meaning as an imperative of justice, designed to protect notably the innocent, but also the culprit, in their legitimate interest to defend themselves on grounds of the principle *in dubio pro reo*. Hence, it should not be blurred by assimilating the standards of scientific validity required from both parties for the evidence they proffer, respectively.

[23] Notwithstanding their somewhat condescending attitude toward cognitive capacities of jurors, let alone judges, with regard to certain scientific niceties that all of the critics seem to understand without difficulty

[24] Notwithstanding the fact that most common law legal systems incorporate cross-examinations to utilize the dynamics of an adversarial interaction between prosecution and counsel in order to scrutinize the evidentiary material. This adversarial aspect has a purely auxiliary (instrumental) sense; it is not an expression of the genuine character of the trial – not more so, at any rate, than former (now fortunately discharged) methods of ascertaining matters of fact, such as torture. Its purpose is to effectively serve the primary goals of the trial as indicated above. It is not itself one of these goals but a method to achieve them.

Summing all this up, from the perspective of the prosecution and for the aim of conviction, expert testimony based on fMRI lie detection is useless and thus inadmissible. From the perspective of the defense counsel, however, the very same method of evidence may well be suitable and admissible. “Slight support (or weak evidence),” as Schauer puts it, “ought not to be good enough for scientists, but it is often sufficient for the law” (Schauer 2010, p. 1208). And here all the more so, since informal, clandestine, unprofessional, unreliable, and uncontrollable methods of assessing a defendant’s or a witness’ credibility in their testimony are common practice and entirely unavoidable in criminal trials. They include judgments on factors such as “whether a witness looks up or down, fidgets, speaks slowly or quickly, and speaks with apparent confidence” (Schauer 2010, p. 1213), or whether their voices vibrate, they slightly blush or pale, their foreheads show traces of perspiration or their eyelids flicker, etc. The intuitive and entirely unchecked ways of laypersons to interpret such signs as evidentiary clues of veracity or mendacity are certainly much farther off any objective standard of scientific validity than results of fMRI on lie detection. Yet no one could seriously consider their flat-out ban from courtrooms. In short, the admissibility of new methods of evidence on grounds of reliability cannot be assessed entirely independent of what has been admissible from time immemorial.

So the following prognosis appears plausible: Considering the present dynamics of fMRI research on lie detection, its methods will not, and should not, be banned entirely from courtrooms simply for reasons of their alleged infeasibility. Rather, they should be cautiously admitted on a case-by-case basis if proffered by defendants. This proposal, however, stands in need of a range of restrictions and controls, one of them personal, the others factual in character. They elucidate what “cautiously” in the foregoing remark means and requires. I will list them in the following subchapter. Note that we are still concerned here with questions of suitability and admissibility, not yet with problems of a normative in-principle justification of (or objections to) such methods.

Restrictions and Caveats: Personal and Factual

Here’s the personal constraint. As yet, fMRI lie detection appears acceptable only on individuals that have freely consented to its use. None of the known fMRI procedures works to any satisfactory degree if applied against the will of subjects. If they oppose the procedure, they can disrupt its effects by covert countermeasures, i.e., behaviors deployed in order to defy deception detection. Possible countermeasures include physical means, such as biting one’s tongue, or (primarily) mental strategies, such as recalling dramatic and arousing events in one’s life or intensely deflecting one’s attention onto other themes (cf. Ganis et al. 2011; Rosenfeld et al. 2004, on so-called P300-based imagings). With regard to defendants, this simply adds another argument to the one above against an application destined to obtain evidence for their conviction. And as to witnesses, as long as they can avail themselves of such simple ways of making the procedure entirely unreliable,

forced-on fMRI is an inadmissible evidence merely for this reason alone, apart from questions of the legitimacy of such force. However, new methods may arise in the foreseeable future that, with a sufficient degree of reliability, foreclose or restrict subjects' ability to even form mendacious thoughts (see Luber et al. 2009). Such methods would perhaps be apt to even forcibly obtain truthful information from witnesses and would, of course, at the same time pose serious problems of justifiability.[25] At present, however, their practical suitability for the purported aim must be denied.

Now let us take a look at the most obvious and most important factual *caveats*. Some of them have already been touched upon in the arguments of the fMRI skeptics quoted above.

1. Laypeople might, as critics fear, confuse fMRI pictures with something like direct snapshots of the brain during its engagement in cognitive tasks. This is false in various respects, as our above sketch of the scientific principles of fMRI already indicates. Such imagings are computer generated out of thousands of recordings of radio pulses, thus averaging their results over large numbers of single "resonances" of hydrogen protons during extended periods of time and over a range of broadly varying resonant signals. They do not directly show any neural activity (let alone at anyone point in time) but rather make it deducible in a twofold indirect manner: from radio echoes which in turn point to biomarkers that are not part of neural tissue proper but primarily reflect changes in metabolic processes in blood vessels. From these biomarkers, some more steps on the rather extended inferential route from fMRI scans to their courtroom applications still need to be taken. All of these steps require highly skilled professionals in the context of highly sophisticated experimental settings in order to lead to useful insights.
2. Furthermore, up to now by far the most fMRI data obtained in experiments and indicating the presence of either (relatively) normal or abnormal neural processes have their origins in the brains of numerous people, i.e., have been derived from, and averaged over, more or less large numbers of individual subjects in each experiment. This is also true for scans that show brains of people during lying as opposed to others speaking truthfully. The method is designed to "cross out" individual differences that appear on a rather broad scale of variations *within* groups of normal (in our case, truthful) or abnormal (deceptive) people as well as between such groups (Hariri 2009). This makes it difficult, if not impossible, to say with reasonable certitude of any one individual brain scan, which differs in some respects from (loosely speaking) the average "truthful" brain and resembles in just these respects the average "lying" brain, that it actually belongs to the latter group – and thus assures that its owner lied (Faigman 2010; Jones and Shen 2012, p. 356).[26]

[25]Hence, we will briefly return to this topic when discussing in-principle objections to fMRI for lie detection in court (see infra, section "Neuroimaging of Deception (2): Principled Normative Objections?").

[26]To avoid misunderstandings, brains do not lie; people do. The above wording serves as a shortcut only.

3. However, this difficulty of subsuming the results of an individual fMRI scan under a certain type of brain function, and hence of classifying the individual person as belonging to a certain group whose members (say, liars) show, on average, just that type of function, is not the only "translational" problem posed by fMRI. Two more are fairly obvious. One pertains to the highly stylized artificiality of the usually primitive lying tests in the scanner that are actually far off any halfway complex situation of deception in everyday life. And the other, related but more profound, pertains to the fact that people in the test situation are *requested* to lie by the researcher. Hence, their "lies" – if one wants to call such linguistic deviations from what's objectively correct "lies"[27] – are entirely free of stress for those who produce them and, hence, psychologically speaking, do not seem to bear much resemblance to the typical real-world lies (cf. Greely and Illes 2007, pp. 403–404; Editorial Nature Neuroscience 2008). What then do the results of such fMRI "lie test" studies indicate with regard to real-life situations in which the falsehood of a testimony, be it given mendaciously or negligently, may have grave consequences for the life of the individual who utters it? As of yet, no one really knows.
4. When people lie, numerous cortical areas, distributed widely over the whole brain, are involved. Typically, each of these areas is also involved in quite a few other mental activities, possibly very different in character than deceiving. There is no such thing as a "lying area" in the brain.
5. Even the conceptual contours of what is being searched for when a lie-detecting fMRI is applied are not clear. What exactly does "lying" imply? Saying "no" where "yes" would be appropriate? Just telling a somewhat different story from the true one? Omitting something? Concealing, slanting, shading, bending some of what is said? And a fortiori unknown is what kinds of differences in neural activity might correspond to such variations.[28] This difficulty may be overcome to a substantial degree, however, by employing (if suitable for the case in question) "concealed information tests" that do not require any verbal answer on the part of the subject, but only an automatic ("autonomic") response to certain visual stimuli on the part of his brain.

This is a formidable list of possible objections to the use of fMRI for lie detection in criminal trials. However, as we saw above, concerns about validity and reliability of the method can be overcome by first making the requisite distinctions and then, of course, taking all those possible objections into account. With regard to its scientific suitability only, no outright ban is called for.

[27]Which is denied, e.g., by Kanwisher (2009, p. 12). The conceptual question is not very interesting here (and can be decided either way); what is important, however, are possibly profound psychological differences between "lying" in the experimental setting and lying in an important real-world situation.

[28]One must not overlook, however, that these variations on truth and lying are "an omnipresent feature of modern litigation" (Schauer 2010, p. 1194) and thus must be dealt with in court in any case and can only be dealt with under considerable uncertainty, be it with or without the assistance (or distraction) of fMRI.

Neuroimaging of Deception (2): Principled Normative Objections?

There is, however, a further weighty concern, one from a normative perspective. Let's assume that someday fMRI for lie detection will be capable of ascertaining thoughts even from an uncooperative suspect with a high degree of certainty and hence well beyond any of the doubts about its validity discussed above.[29] Would it then be acceptable as evidence? Or would it violate the basic rights of defendants?

Again, we need to draw at least two distinctions in order to tackle this problem: one is between applications on defendants and on witnesses and the other between a compelled use and one consented to by the person concerned.

Compelled Application on Defendants?

Suspects have no legal obligation to actively contribute anything to their conviction. Consequently, they have a right to silence, which includes that no potentially incriminating testimonial evidence be obtained from them against their will or even by force. This privilege against compelled self-incrimination – in its classical Latin form, *nemo tenetur se ipsum accusare* ("no one is obligated to accuse himself") – is a fundamental principle of the law of criminal proceedings. It is also a positive legal norm in many jurisdictions and established in various international covenants.[30] If it comes to precisely determining its scope and content, however, notorious difficulties and long-standing doctrinal controversies arise.[31] They originate mainly from a potential normative conflict with another principle, an equally uncontested norm of procedural criminal law in most jurisdictions: Whereas a suspect must not be forced to actively testify against himself (*nemo tenetur*), his body in all its biological parts may well be examined against his will (if necessary forcibly) and thus be used as physical evidence against him. Put another way, whereas defendants cannot be compelled to reveal anything that might incriminate them with regard to the crime they are suspected of, every piece of their (inner and outer) body may be

[29]This is not sheer science fiction. It may become feasible in the not too distant future by first employing certain measures of brain stimulation that significantly impede a suspect's ability to engage brain networks involved in conscious deceit and then subsequently submit their brains to an fMRI via CIT, i.e., by scanning them for concealed knowledge or memory; cf. Luber et al. (2009).

[30]In the US Constitution, the principle is protected in the Fifth Amendment. The International Covenant on Civil and Political Rights (1966) expressly warrants it in Art. 14 para. 3 (g). And it is usually derived from Art. 6 para. 2 of the European Convention on Human Rights of 1950 ("presumption of innocence until proven guilty") – not a logical deduction, to be sure, but a plausible normative derivation.

[31]The arguments in these controversies are probably similar in most jurisdictions where the principle is guaranteed; they certainly are, for instance, in the US and the German scholarly debate; of the vast doctrinal literature in both these countries, cf. only Pardo (2008), Fox (2009), Verrel (2001), Eidam (2007) – each with a long list of further sources of reference.

forcibly scrutinized to reveal whatever may indicate their involvement in that crime. Obviously, these two principles may collide with each other time and again. Thus, they pose the problem of how to demarcate their respective normative purviews.

In decades of jurisprudential debate about this problem, a variety of differentiating criteria have been developed by courts and scholars, but none has proven entirely convincing or capable of achieving unanimous consent. In the US judicature (as an example for a common-law jurisdiction), the distinction between what is forbidden by the Fifth Amendment and what is not hinges on whether the compelled incriminating evidence from a suspect is "testimonial" or "physical," the former being unconstitutional whereas the latter doubtlessly legitimate (cf. Schmerber v. California 1966).[32] The respective distinction in the judicature of the German Federal Criminal Court (as exemplary for continental or "civil law" systems) rather turns on whether the suspect is compelled to *actively* contribute incriminating evidence of whatever kind and amount (forbidden) or only forced to *passively* endure bodily examinations of whatever kind or scale (allowed).[33]

Against the backdrop of these judicial criteria – and, in fact, of most other criteria proposed in scholarly discourse – the question whether fMRI scans for lie detection would violate the right to silence (*nemo tenetur*, Fifth Amendment) raises a puzzling problem. Is (forcibly) scrutinizing a suspect's brain in order to detect (via a series of inferences) certain mental entities or processes in their consciousness "physical" or "testimonial" in the sense deployed by US courts? That is, does it amount to nothing more than attaining information about physiological processes in one of their organs (their brain) – which would clearly be allowed and hence legitimate ground for whatever (perhaps incriminatory) conclusion to be drawn from it? Or is it rather akin to making them disclose (via their brains) knowledge, thoughts, and memories from their innermost mental sphere – which would just as clearly be forbidden by the *nemo tenetur* principle, or the Fifth Amendment for that matter? Or from the perspective of the German law, is what is extorted from the suspect an *active* contribution to their potential conviction (namely, cascades of brain activities) – which would be forbidden? Or are they only compelled to passively bear an examination of one of their physical organs (from which observers may then draw their own conclusions) – which would clearly be allowed?

In a sense fMRI for lie detection is simply both, depending on how one looks at it. It promises "distinctly testimonial-like information about a person's mind that is packaged in demonstrably physical-like form" (Fox 2009, p. 792). Epistemologically, such perspectivism, yielding different descriptions for the same object, is entirely unproblematic. Normatively speaking, however, fMRI cannot be both allowed and forbidden at the same time. So we must decide what we take it for.

[32]For a thorough examination of the "testimonial/physical" distinction and convincing critique of its shortcomings (see Fox 2009; see also Pardo 2008).

[33]It goes without saying that such an examination must not significantly threaten the suspect's health. For a largely convincing critique of this "active vs. passive" criterion see Verrel (2001).

This is a genuinely normative problem. It cannot be resolved by comparing phenomenal similarities and dissimilarities between fMRI and paradigm types of testimonial as opposed to physical evidence (or of actively contributing vs. passively bearing on the part of the suspect). Instead, we must clarify what deeper principle a suspect's right to silence is based on and then ask whether or not a compelled fMRI for lie detection would contravene that principle's basic normative sense. Simply, albeit correctly, ascertaining that compelled testimonial evidence (or compelled active self-incrimination) is unlawful in criminal proceedings does not explain why this is and should continue to be so and hence does not help us solve our problem.

We cannot enter here the labyrinth of intricate arguments that have been put forward to cope with this problem.[34] Suffice it to say the following: What decisively matters for the prohibition of forcing suspects to testify (actively and/or "testimonially") against themselves is the fact that such compulsion seizes the authority of control over thoughts, memories, knowledge, and other mental processes – in short, the inventory of one's mind and thus over the core of one's personality.[35] Exerting external control over someone's mind in such a way, be it by compulsive threats or by irresistible physical force, deprives a person of a constitutive element of personhood at large. That is to say, being a person in the full normative sense of the concept does not merely involve having a mental inventory of thoughts, reminiscences, emotions, etc. (all of which certainly constitute the individual "I") but also being in immediate command of the processes that dispose of and deal with such elements of the inner self (to the extent they are at one's willful disposal at all).

This is an evaluative, not a descriptive statement, and thus not accessible to scientific proof or refutation. It does, however, plausibly fit our concept of personhood. And it is also significantly confirmed by the historic background against which it should be judged: the so-called "inquisitional" type of criminal trial which for centuries was characterized by outright barbarous procedures of coercing suspects to confess, viz., by subjecting them to torture. The principle of *nemo tenetur* is designed to ban from criminal trials not only such methods but also the related goal of seizing control over a suspect's mind by gaining compelled access to their thoughts and thus depersonalizing them in a certain way and for a certain purpose. Hence, the innermost control over one's mind is declared legally sacrosanct from unwanted access and use by the state in criminal proceedings.

This provides us with a clear answer to our question: Compulsory fMRI for lie detection in suspects is illegitimate and excluded from criminal trials. For in such cases, the brain would not be searched to ascertain physiological activities, which in themselves are uninteresting for any evidentiary purpose, but to gain access to the corresponding mental processes by displacing the defendant as the subject of

[34]For exemplary scholarly analyses, see sources cited supra, n. 31.

[35]A very similar conception in Fox (2009, 796 pp.); relatedly for the German law Verrel (2001, 246 pp.).

control over these processes and their public proclamation. It is of no significance whether this happens via compulsion or via circumventing the controlling position of the individual altogether, i.e., by usurping their privileged access to their own mental sphere and thus extracting part of its content, as it were done in compulsory fMRI. This differentiation between neural activities and correlated mental processes does not rely on a philosophically flawed strong (or "substance") dualism of a Cartesian provenance, as Fox worries (Fox 2009, 793pp.). All monistic conceptions of the mind-brain relationship also presuppose a certain correlation (not interaction!) between mental states and their neural underpinnings, be it one of possible "reduction" of mind to brain processes, one of different "aspects" of one and the same entity, or any of the other conceptions proposed to grasp this specific correlational setting.[36] In our context, we may safely ignore these contested metaphysical problems, anyway. For our distinction between mental processes and brain activities is a purely normative one. Its adequacy does not depend on any particular philosophical position on the mind-brain problem.

Compelled Application on Witnesses?

Witnesses, of course, do not regularly have a right to refuse testimony (specific exceptions aside), much less a right to lie while giving evidence. Would it be legitimate to subject them to fMRI lie detection against their will? As long as they can evade the goal of the procedure by employing simple and effective countermeasures, its application is useless and thus inadmissible.[37] Scientific progress may, however, develop fMRI methods – perhaps, as the case may be, in conjunction with certain forms of brain stimulation (Luber et al 2009) – which are largely immune to such countermeasures. Then? There is no principled or (depending on the jurisdiction in question) constitutional objection that would decisively rule out such an application per se. However, in liberal states or (in Rawlsian terms) in well-ordered societies, the methods of compulsion must, of course, be restricted. Such states should not resort to physical force in order to coerce witnesses into brain scanners. And with regard to our above conclusion that controlling the access to one's own mind is a constitutive function of personhood, it is decidedly preferable for liberal states to also avoid other (nonphysical) forms of compelling fMRI for lie detection on disinclined witnesses.

[36]With, perhaps, the exception of an "eliminative monism" that attempts to somehow dispose of the mental side of brain processes altogether – not a very attractive philosophical position. It is unpromising to deny that we do subjectively experience mental events such as phenomenal states (of "what it's like"), even if they are nothing but "the other side" of brain processes (which we do not subjectively experience).

[37]As we saw above (II.3, supra), if they consent to the fMRI and want to testify in favor of the defendant's innocence, the procedure is not entirely unfeasible and hence may well be considered admissible.

fMRI for Lie Detection on Freely Consenting Suspects or Witnesses Who Request It?

There are no principled objections against such applications in criminal proceedings. And there is no reason to worry about an undue "functionalization" of persons and hence perhaps a violation of their dignity[38] through an fMRI they consented to. Their dignity begins with their choice. As to questions of admissibility, which are a different matter, we may now refer to our above arguments for a differential solution with respect to the different roles and the respective onus of proof of the parties to a criminal trial (see section "Neuroimaging of Deception: Feasibility? Admissibility?").

Conclusions: fMRI for Lie Detection

With regard to the admissibility and legitimacy of fMRI for lie detection, four conclusions from our foregoing considerations suggest themselves:

1. There should not be an outright ban, not even a present-day moratorium, on fMRI-based veracity tests in criminal trials. For prosecutorial purposes of conviction, however, the method is entirely unsuited and hence inadmissible. Not so, by contrast, for the much more modest goal of exonerating the accused under the principle of *in dubio pro reo* ("benefit of the doubt"). That the probative value of fMRI imaging will approach anything near certitude (comparable to DNA tests) in the foreseeable future is extremely unlikely. On the other hand, it is way above any dubious hyperboles of "junk science." Rather, it is an objective of serious research, and it will, in all likelihood, not stagnate at its current level. That this level should make fMRI for lie detection in forensic settings entirely unfeasible, i.e., even for the purposes of defendants and their counsel, is unconvincing.
2. However, the efficiency of fMRI tests can as yet rather easily be undermined by destructive countermeasures of unwilling subjects. Hence, they are only feasible, and thus admissible, if applied on freely consenting defendants or witnesses.
3. On the part of defendants, this holds also for the basic normative reason of their right to silence (principle of *nemo tenetur*) as protected in various international covenants and in national constitutions such as the Fifth Amendment in the United States. This principle forbids extracting possibly incriminating testimonial knowledge from suspects against their will by circumventing their personal control over their own thoughts and memories. This principle does not apply to witnesses. However, because controlling the access to one's own mind is a constitutive element of personhood, witnesses should not be compelled to undergo fMRI for lie detection either. If, by contrast, suspects or witnesses

[38] As was erroneously done in an early decision of the German Federal Criminal Court with regard to polygraphy that was requested by the defendant; see BGHSt 5, 332 (1954).

consent to, or even require, an fMRI test in order to strengthen their credibility, no principled objection stands in the way of complying with their request.

4. If fMRI results are admitted, it must be pointed out to jurors or judges by qualified experts that the value of fMRI as circumstantial evidence to assess the truthfulness of a testimony is usually low to marginal. This holds even for the modest purpose of exonerating the defendant under the principle *in dubio pro reo*. The reasons for this limited probative value must be elucidated to the triers. In particular, the above-listed caveats (see section "Restrictions and Caveats: Personal and Factual" (1–5)) should be expounded in order to counter whatever "seductive allure" or "Christmas tree" effect fMRI results might potentially exert.

FMRI for "Neuroprediction": Assessing Future Dangerousness

The necessity to make forensic prognoses about the future dangerousness of criminal defendants may occur in (at least) two different contexts: for most common law jurisdictions in the sentencing phase of a trial and furthermore (and in "continental" systems *only*) in procedures of preventive detention, which include the coerced confinement of violent sexual offenders as it is established in the so-called sexual predator statutes in many US jurisdictions (cf. Nadelhoffer et al. 2012, 75pp.). Could fMRI be employed for such risk assessments on potentially dangerous people?

Here is the premise that any sensible answer must reckon with: What is at stake in both these types of forensic predictions is the option for the state to impose sanctions on someone for something they have not done (but are only feared to eventually do in the future). It is clear at once that such a practice is somewhat of a borderline case for any legitimate legal order committed to principles of justice.[39] Against this background, it seems cogent that the state is obliged to utilize and exhaust all available, scientifically acceptable methods to ascertain the prognosis of the future dangerousness of a delinquent, given that such a prognosis is indispensable once the question of dangerousness has seriously arisen in a case. At present, it is based on the expertise of (usually two) psychiatrists, a cognitive basis fraught with uncertainties and all too often error (cf. Ennis and Litwack 1974; Thornberry and Jacoby 1979; Monahan et al. 2001) – or even on nothing but juries' or judges' intuitions, a base that's patently unfit to master a task of such importance (cf. Reidy et al. 2013).

[39]This throws into sharp relief concerns of justice about the established practice in many common law jurisdictions to impose harsher penalties than are matched by the degree of guilt (or viciousness, as it were) realized by someone's crime because the perpetrators are believed to pose a future danger to society. One should not be *punished* for something one has not committed (and hence should also not suffer an *extra* penalty beyond what their crimes alone make them deserve) – though one may certainly be *kept detained* if one is rightly assessed to pose a substantial risk of seriously harming others in the future.

Might fMRI scans provide adjuvant support here? Currently, there are at least two mental (or, for that matter, neural) predispositions to re-offend for which fMRI results might yield a sufficiently validated diagnostic and hence prognostic basis: pedophilia and psychopathy. Rather strong empirical evidence indicates a significantly higher risk of recidivism in perpetrators with one (let alone both) of these two mental conditions.[40] It seems that both these conditions can already be assessed with a sufficient degree of reliability by fMRI (for pedophilia cf. Ponseti et al. 2012; Wiebking et al. 2012; for psychopathy Wahlund and Kristiansson 2009; Anderson and Kiehl 2013). In light of these findings, states are not only entitled, but do indeed have an obligation to include "neuropredictive" methods in the prognostic procedures underlying the imposition of measures of preventive detention. Given that the question of dangerousness cannot be avoided, given furthermore how much is at stake for the person concerned (indefinite confinement for something they haven't done), legal decision makers certainly "need to be equipped with the best possible predictions concerning future dangerousness" (Nadelhoffer et al. 2012, p. 76).

Clear as this is, it still raises a few qualms and calls for a few caveats. Not among these qualms, by the way, are concerns about possible violations of the "*nemo tenetur*" principle (in the United States, the Fifth Amendment). That principle only grants protection against self-incriminating testimony with regard to a crime one is suspected to have committed. It does not, however, preclude any effort to obtain as many clues as possible (by legally approved means, of course) concerning their future dangerousness.[41] Here are three caveats:

1. Obviously, fMRI-based predictions cannot replace the classical psychiatric methods of risk assessment. They can only provide an additional or, as the case may be, complementary source of cognition in order to ascertain the broadest base feasible for predictions of potential risks of criminal recidivism.
2. It must be clearly pointed out what fMRI scans can and what they cannot demonstrate. That someone has pedophilic inclinations does not necessarily mean that they will in fact encroach upon children's sexual integrity. According to recent empirical research, quite a few more men (and even women) than ever commit a pedophilic crime do in fact have such inclinations but are able to restrain themselves vis-à-vis the threat of legal punishment (Wurtele et al. 2013). Thus, knowing that someone is a pedophile does not include foreknowing that they will sexually assault children, even if they have already committed at least one such assault in the past (as is regularly presupposed in cases of preventive detention). The same argument holds, and perhaps even more so, for violent offenders who are diagnosed to be psychopaths.

[40]For pedophiles, (see Wilson et al. 2011; Hanson and Morton-Bourgon 2005); for psychopathic offenders (see Olver and Wong 2006; Rice and Harris 2013); for psychopathic sex offenders (see Porter et al. 2009).

[41]The rather unfortunate fact notwithstanding that many common-law jurisdictions confound grounds for punishing past deeds with grounds for predicting future dangerousness (cf. supra, n. 39).

3. On the other hand, the empirical research mentioned above indicates a significant higher risk of future sexual assaults in people with pedophilic appetence than with other sexual, e.g., homo- or heterosexual, orientations. Furthermore, by far not all who sexually attack children are genuine pedophiles; some simply take advantage of the defenselessness of children, without a specific or exclusive sexual drive toward their young victims. Hence, being able to demonstrate with a high degree of reliability that someone who committed a sexual assault on a child really is pedophilic, as fMRI scans are apparently capable of, actually provides an important element for a sufficient cognitive basis to assess their risk of recidivism. And again, the same also holds for the prediction of future dangerousness of psychopathic violent offenders.

This warrants the following prospect: fMRI (and other) brain scans are or, at any rate, will certainly be well suited to contribute valuable assistance to the difficult task of prognosticating criminal recidivism in certain types of sexual or violent offenders. Thus, they might help us clear up some of the dark spots in the practice of preventive detention. We must, however, be careful not to allow them to rather add one further spot to that record: an excessive trust and hence a kind of mechanistic application of their results, based on an overassessment of their capabilities and an underestimation of their limits. There is no such thing as a "criminal brain." There are, however, mental dispositions to act in certain ways that raise the risk of becoming criminal in their possessors. Like all dispositions to act, they have their proximate (though of course not their only) causal source in people's brains. To identify them there and to draw legally relevant conclusions with the necessary skeptical diligence – that is the future task of courts, psychiatrists, and legal scholars. Probably rather sooner than later, the rapidly developing methods of neuroimaging will provide indispensable support.

Cross-References

▶ Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response
▶ Neuroimaging Neuroethics: Introduction
▶ Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges

References

Aggarwal, N. K. (2009). Neuroimaging, culture, and forensic psychiatry. *Journal of the American Academy of Psychiatry and the Law, 37*(2), 239–44.

Aguirre, G. K., & D'Esposito, M. (1999). Experimental design for brain fMRI. In C. T. W. Moonen & P. A. Bandettini (Eds.), *Functional MRI* (pp. 369–380). New York: Springer.

Anderson, N. E., & Kiehl, K. (2013). Functional neuroimaging and psychopathy. In K. A. Kiehl & W. Sinnott-Armstrong (Eds.), *Handbook on psychopathy and law* (pp. 131–149). Oxford: Oxford University Press.

Aronson, J. D. (2010). The law's use of brain evidence. *Annual Review of Law and Social Science, 6*, 93–108.

Ben-Shakar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology, 88*(1), 131–151.

Best v. Lowe's Home Centers, Inc. (2009). 563 F.3d 171 (6th Cir.).

BGHSt = Entscheidungen des Bundesgerichtshofs in Strafsachen, amtliche Sammlung (Decisions of the Federal Criminal Court of Germany, official collection).

Bigler, S. E., Allen, M., & Stimac, G. K. (2012). MRI and functional MRI. In J. R. Simpson (Ed.), *Neuroimaging in forensic psychiatry* (pp. 27–40). Chichester: Wiley.

Brown, T., & Murphy, M. (2010). Through a scanner darkly: Functional neuroimaging as evidence of a defendant's past mental states. *Stanford Law Review, 62*, 1119–1208.

Daubert v. Merrill Dow Pharmaceuticals Inc. (1993). *509 U.S. 579*.

Editorial (2008). Deceiving the law. *Nature Neuroscience 11/11*, 1231.

Eggen, J. M., & Laury, E. J. (2012). Toward a neuroscience model of tort law: How functional neuroimaging will transform tort doctrine. *Columbia Science and Technology Law Review, 13*, 235–306.

Eidam, L. (2007). *Die strafprozessuale Belastungsfreiheit am Beginn des 21. Jahrhunderts*. Frankfurt/Berlin: Peter Lang.

Ennis, B. J., & Litwack, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review 62*, 693–752.

Faigman, D. L. (2010). Evidentiary incommensurability. A preliminary exploration of the problem of reasoning from general scientific data to individualized legal decision-making. *Brooklyn Law Review 75*, 1115–1136.

Feigenson, N. (2006). Brain imaging and courtroom evidence: On the admissibility and persuasiveness of fMRI. *International Journal of Law in Context, 2*(3), 233–255.

Fox, D. (2009). The right to silence as protecting mental control. *Akron Law Review, 42*, 763–801.

Gamer, M., Bauermann, T., Stoeter, P., & Vossel, G. (2007). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human Brain Mapping, 28*, 1287–1301.

Gamer, M., Klimecki, O., Bauermann, T., Stoeter, P., & Vossel, G. (2012). fMRI-activation patterns in the detection of concealed information rely on memory-related effects. *Social Cognitive and Affective Neuroscience, 7*, 506–515.

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage, 55*, 312–319.

Gary Smith v. State of Maryland. (2012). http://mdcourts.gov/opinions/coa/2011/10a11.pdf. Accessed 17 Dec 2013.

Giridharadas, A. (2008, September 15). India's novel use of brain scans in courts is debated. *New York Times*.

Granacher, R. C. (2012). Potential uses of neuroimaging in personal injury civil cases. In J. R. Simpson (Ed.), *Neuroimaging in forensic psychiatry* (pp. 201–213). Chichester: Wiley.

Greely, H. T., & Illes, J. (2007). Neuroscience-based lie detection: The urgent need for regulation. *American Journal of Law & Medicine, 33*, 377–431.

Hakun, J. G., Seelig, D., Ruparel, K., Loughead, J. W., Busch, E., Gur, R. C., & Langleben, D. D. (2008). fMRI investigation of the cognitive structure of the concealed information test. *Neurocase, 14*, 59–67.

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*, 1154–1163.

Hariri, A. R. (2009). The neurobiology of individual differences in complex behavioral traits. *Annual Review of Neuroscience, 32*, 225–247.

Jones, O., & Shen, F. (2012). Law and neuroscience in the United States. In T. M. Spranger (Ed.), *International neurolaw: A comparative analysis* (pp. 349–380). Berlin/Heidelberg: Springer.

Jones, O. D., Buckholtz, J. W., Schall, J. D., & Marois, R. (2009). Brain imaging for legal thinkers: A guide for the perplexed. *Stanford Technology Law Review, 5*.

Kanwisher, N. (2009). The use of fMRI in lie detection: What has been shown and what has not. In E. Bizzi, S. E. Hyman, M. E. Raichle, N. Kanwisher, E. A. Phelps, S. J. Morse, W. Sinnott-Armstrong, J. S. Rakoff, H. T. Greely, et al. (Eds.), *Using imaging to identify deceit: Scientific and ethical questions* (pp. 7–13). Cambridge, MA: American Academy of Arts and Sciences.

Langleben, D. D., Willard, D. F. X., Moriarty, J. C. (2012). Brain Imaging of Deception. In J. R. Simpson (Ed.), *Neuroimaging in forensic psychiatry* (pp. 217–236). Chichester: Wiley.

Langleben, D. D., & Moriarty, J. C. (2013). Using brain imaging for lie detection: Where science, law, and policy collide. *Psychology, Public Policy, and Law, 19*(2), 222–234.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*, 869–878.

Logothetis, N. K., & Pfeuffer, J. (2004). On the nature of the BOLD fMRI contrast mechanism. *Magnetic Resonance Imaging, 22*, 1517–1531.

Luber, B., Kinnunen, L. H., Rakitin, B. C., Ellsasser, R., Stern, Y., & Lisanby, S. H. (2007). Facilitation of performance in a working memory task with rTMS stimulation of the precuneus: Frequency and time-dependent effects. *Brain Research, 1128*, 120–129.

Luber, B., Fisher, C., Appelbaum, P. S., Ploessner, M., & Lisanby, S. H. (2009). Non-invasive brain stimulation in the detection of deception: Scientific challenges and ethical consequences. *Behavioral Sciences and the Law, 27*, 191–208.

MacLaren, V. V. (2001). A qualitative review of the guilty knowledge test. *Journal of Applied Psychology, 86*(4), 674–683.

Matthews, P. M., & Jezzard, P. (2004). Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery and Psychiatry, 75*, 6–12.

McLaughlin, B., & Bennett, K. (2013). Supervenience. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/supervenience/. Accessed 15 Dec 2013.

Merkel, R. (2008). *Willensfreiheit und rechtliche Schuld. Eine strafrechtsphilosophische Untersuchung*. Baden-Baden: Nomos.

Merkel, R. (2011). Schuld, Charakter und normative Ansprechbarkeit. In M. Heinrich & C. Jäger (Eds.), *Strafrecht als Scientia Universalis, Festschrift für Claus Roxin zum 80* (Geburtstag, Vol. I, pp. 737–761). Berlin: De Gruyter.

Mobbs, D., Lau, H. C., Jones, O. D., & Frith, C. D. (2007). Law, responsibility, and the brain. *PLOS Biology, 5*, 693–700.

Monahan, J., Steadman, H. J, Silver, E., Appelbaum, B. S., Clark Robbins, P., Mulvey, E. P., Roth, L. H., Grisso, T., & Banks, S. (2001). *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*. Oxford: Oxford University Press.

Moriarty, J. C. (2009). Visions of deception: Neuroimages and the search for truth. *Akron Law Review, 42*, 739–762.

Moriarty, J. C., Langleben, D. D., & Provenzale, J. M. (2013). Brain trauma, PET scans and forensic complexity. *Behavioral Sciences and the Law, 31*(6), 702–720.

Morse, S. J. (2006). Brain overclaim syndrome and criminal responsibility: A diagnostic note. *Ohio State Journal of Criminal Law, 3*, 397–412.

Morse, S. J. (2012). Neuroimaging evidence in law: A plea for modesty and relevance. In J. R. Simpson (Ed.), *Neuroimaging in forensic psychiatry* (pp. 341–357). Chichester: Wiley.

Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K., Mansfield, A., Sinnott-Armstrong, W., & Gazzaniga, M. (2012). Neuroprediction, violence, and the law: Setting the stage. *Neuroethics, 5*, 67–99.

National Research Council of the National Academies. (2003). *The polygraph and lie detection*. Washington D.C.: The National Academies Press.

Olver, M. E., & Wong, S. C. P. (2006). Psychopathy, sexual deviance, and recidivism among sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 18*, 65–82.

Oullier, O. (2011). Clear up this fuzzy thinking on brain scans. *Nature, 483*, 7.

Pardo, M. S. (2008). Self-incrimination and the epistemology of testimony. *Cardozo Law Review, 30*, 1023–1046.

Ponseti, J., Granert, O., Jansen, O., Wolff, S., Beier, K., Neutze, J., Deuschl, G., Mehdorn, H., Siebner, H., & Bosinski, H. (2012). Assessment of pedophilia using hemodynamic brain response to sexual stimuli. *Archives of General Psychiatry, 69*, 187–194.

Porter, S., ten Brinke, L., & Wilson, K. (2009). Crime profiles and conditional release performance of psychopathic and non-psychopathic sexual offenders. *Legal and Criminological Psychology, 14*, 109–118.

Pustilnik, A. (2009). Violence on the brain: A critique of neuroscience in criminal law. *Wake Forest Law Review, 44*, 183–237.

Putzke, H., Scheinfeld, J., Klein, G., & Undeutsch, U. (2009). Polygraphische Untersuchungen im Strafprozess. *Zeitschrift für die gesamte Strafrechtswissenschaft, 121*, 607–644.

Raichle, M. E. (2009). An introduction to functional brain imaging in the context of lie detection. In E. Bizzi, S. E. Hyman, M. E. Raichle, N. Kanwisher, E. A. Phelps, S. J. Morse, W. Sinnott-Armstrong, J. S. Rakoff, H. T. Greely, et al. (Eds.), *Using imaging to identify deceit: Scientific and ethical questions* (pp. 3–6). Cambridge, MA: American Academy of Arts and Sciences.

Reidy, T. J., Sorensen, J. R., & Cunningham, M. D. (2013). Probability of criminal acts of violence: A test of jury predictive accuracy. *Behavioral Sciences and the Law, 31*, 286–305.

Rice, M., & Harris, G. T. (2013). Psychopathy and violent recidivism. In K. A. Kiehl & W. Sinnott-Armstrong (Eds.), *Handbook on psychopathy and law* (pp. 231–249). New York: Oxford University Press.

Robinson, H. (2013). Dualism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/dualism/. Accessed 15 Dec 2013.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology, 41*, 205–219.

Schauer, F. (2010). Can bad science be good evidence? Neuroscience, lie detection, and beyond. *Cornell Law Review, 95*, 1191–1219.

Schmerber v. California. (1966). *384 U.S. 757*.

Seaman, J. (2009). Black Boxes: fMRI lie detection and the role of the jury. *Akron Law Review, 42*, 931–939.

Sinnott-Armstrong, W., Roskies, A., Brown, T., & Murphy, E. (2008). Brain images as legal evidence. *Episteme, 5*, 359–373.

Sip, K. E., Roepstorff, A., McGregor, W., & Frith, C. (2007). Detecting deception: The scope and limits. *Trends in Cognitive Science, 12*(2), 48–53.

Smart, J. J. C. (2013). Mind/brain identity theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/mind-identity/. Accessed 15 Dec 2013.

Spence, S., Kaylor-Hughes, C. J., Brook, M. L., Lankappa, S. T., & Wilkinson, I. D. (2008). ‘Munchausen’s syndrome by proxy’ or a ‘miscarriage of justice’? An initial application of functional neuroimaging to the question of guilt versus innocence. *European Psychiatry, 23*, 309–314.

Taylor, K. (2012). *The brain supremacy*. Oxford: Oxford University Press.

Thornberry, T. B. & Jacoby, J. E. (1979). *The criminally insane: A community follow-up of mentally ill offenders*. Chicago: University of Chicago Press.

United States v. Semrau, US Court of Appeals, 6th Cir. (2012). http://www.ca6.uscourts.gov/opinions.pdf/12a0312p-06.pdf. Accessed 17 Dec 2013.

Uttal, W. (2009). *Neuroscience in the courtroom. What every lawyer should know about the mind and the brain*. Tucson: Lawyers & Judges Publishing.

Verrel, T. (2001). *Die Selbstbelastungsfreiheit im Strafverfahren*. München: C.H. Beck.

Vincent, N. (2011). Neuroimaging and responsibility assessments. *Neuroethics, 4*, 35–49.

Wahlund, K., & Kristiansson, M. (2009). Aggression, psychopathy and brain imaging - Review and future recommendations. *International Journal of Law and Psychiatry, 32*, 266–271.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*, 470–477.

Wiebking, C., Sartorius, A., Dressing, H., & Northoff, G. (2012). Pedophilia. In J. R. Simpson (Ed.), *Neuroimaging in forensic psychiatry* (pp. 99–111). Chichester: Wiley.

Wilson, R. J., Abracen, J., Looman, J., Picheca, J. E., & Ferguson, M. (2011). Pedophilia: An evaluation of diagnostic and risk prediction methods. *Sexual Abuse: A Journal of Research and Treatment, 23*(2), 260–274.

Wurtele, S. K., Simons, D. A., & Moreno, T. (2013). Sexual interest in children among an online sample of men and women: Prevalence and correlates. *Sexual Abuse: A Journal of Research and Treatment*. Online first: http://sax.sagepub.com/content/early/2013/11/11/1079063213503688.full.pdf+html (last visited 10 January 2014).

Responsibility Enhancement and the Law of Negligence

85

Imogen Goold and Hannah Maslen

Contents

Abstract

Much of the law is concerned with ascribing responsibility. The law of negligence looks for the person who acted without due care and places on them the responsibility for the outcome. The criminal law responds punitively to those who break its rules, but the accused can avoid being held wholly or partially responsible if she can point to evidence that showed she lacked the capacity to be in charge of her actions. Contract law is based around competent individuals voluntarily taking on

I. Goold (✉)
Faculty of Law, University of Oxford, St Anne's College, Oxford, UK
e-mail: Imogen.goold@law.ox.ac.uk; imogen.goold@st-annes.ox.ac.uk

H. Maslen
Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK
e-mail: Hannah.maslen@new.ox.ac.uk; hannah.maslen@philosophy.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_168,

obligations to one another and holds them responsible if they fail in them. Responsibility is demonstrably a key concept in the law of England, and therefore we should be particularly interested in technology that can affect an individual's capacity to be responsible. This chapter looks at one particular area of English law in which new drugs are potentially changing our capacities and hence (arguably) our responsibilities: cognitive enhancers. Using a scenario involving a tired surgeon, this chapter briefly examines the question: When, if at all, might someone be legally obliged to take a cognitive-enhancing drug?

Introduction

As the use of cognitive enhancers increases, so does the potential for legal claims. As the technology currently stands, the most likely claims will be for harm resulting due to negligence, and most often in the context of driving and provision of professional services, particularly clinical practice. This is in part because, at present, the most widely used cognitive-enhancing pharmaceuticals can be used to combat decreases in wakefulness and cognitive capacity that arise due to fatigue (Baranski et al. 2002; Caldwell 2005; Grady et al. 2010; Hart et al. 2006; Sugden et al. 2012). Consequently, in situations where the injury complained of happened due to the defendant's fatigue, it might be open to claim that the defendant could have avoided this if she had taken a cognitive enhancer. From a legal perspective, this raises three questions: Is there a duty to take such an enhancer (liability for omissions)? Does failing to take an enhancer breach the standard of care owed to the claimant? Can it be said that the failure to take the enhancer *caused* the injury? Each of these questions is briefly examined using a scenario involving a tired surgeon. The discussion focuses on the use of modafinil, as there is growing evidence of its efficacy in addressing fatigue. These questions, and most of the conclusions drawn in response to them, could be extrapolated to other contexts.

An alternative liability situation arises if the defendant *has* taken the enhancer and the claimant alleges that doing was negligent and caused the injury he suffered. Here, the questions would be: Does taking the enhancer breach the standard of care owed to the claimant? Can it be said that taking the enhancer *caused* the injury? These questions are legally simpler than those relating to a failure to take the enhancer, and their answers will be essentially determined by the possibility of linking taking the enhancer with causing harm, which is a matter of evidence more than law. For these reasons, these questions are considered only briefly at the end of this chapter.

There is also evidence that cognitive enhancers can lift cognitive capacity above the norm – either one's own norm or the general norm. For example, it appears that for some people, modafinil has the effect of increasing the length of time for which they can remain focused on a task, and for some it has the effect of increasing concentration and suppressing the need for sleep. In a situation where a person's normal capacity to focus would naturally wane and potentially lead to injurious

errors, taking a cognitive enhancer might prevent such errors. By increasing the person's capacity to concentrate beyond the norm, it might then be argued that their level of responsibility should rise accordingly. Such an approach has been explored in the philosophical literature and is briefly touched on below, but for the most part, the English legal system treats responsibility in negligence as a threshold concept. Given this, it is highly unlikely that English law will incorporate a system of enhanced responsibility in the near future, and so this issue is not dwelt upon here. However, it is noted that as our ability to cognitively enhance improves, this is likely to become an important area of legal research.

The chapter concludes by drawing together the threads of each of these three sections to argue that while there is support in the ethics literature for obliging the use of enhancers, and suggestion that responsibility should track capacity, it is not likely that these calls will be heeded by the English courts or legislature in the near future. On this point, the chapter briefly examines the relationship between legal and ethical obligations, as well as the impact of public demand, to demonstrate why these three sources of direction about how we should act do not always map onto one another.

Calls for Legal Requirements to Take Enhancers

The suggestion that individuals in certain jobs should take cognitive enhancers has emerged from a variety of places. Philosophers, lawyers, stakeholders, and even the professionals around whom the debate centers have considered whether a duty to enhance will be likely and/or justified. Some writers focus on the idea that it would be a good idea practically – it would be prudent to use enhancers to make things safer if possible. Others have suggested that there could be a moral obligation for professionals in high-risk jobs to enhance. Still others have described how changing attitudes and practices might indirectly generate a legal duty for these professionals to enhance. The recent report from the Joint Academies' workshop on "Human Enhancement and the Future of Work" (2012) suggests that people may see a moral obligation for some individuals to use enhancers at work or even demand that they do so:

> [O]ccupations that require particular patterns of focus could benefit from enhancements that facilitate achieving such patterns. For example, surgeons may need to be able to concentrate for extended periods, whereas other jobs such as air traffic control can require very rapid reactions during periods of relative uniformity. As an extrapolation to this, it is possible that in these high-responsibility occupations enhancement could be seen as a moral obligation, or even demanded by the public.

While a perceived moral obligation would make the conduct of the high-responsibility professionals open to moral evaluation, public demands for enhancement could put pressure on those guiding professional practice to impose a duty to enhance. Ethicists have suggested that there are other domains in which enhancement should be routinely adopted. Discussing cognitive enhancement in courts, Sandberg et al. (2011) claim:

> As our range and knowledge of cognitive enhancers increases, some enhancers should be provided, as we now provide coffee, if they are safe and do not bias judgement. Substances like modafinil may well meet this criterion. Given the stakes, we have a moral imperative to investigate ways to utilize these new technologies to improve cognitive performance in the courtroom. Innocent people's lives may well depend on them.

The suggestion that there may be a moral obligation to enhance is not confined to academics. Some of the professionals who might be expected to take enhancers also share the view. The editors of the Mayo Journal of Clinical Proceeding, both medical doctors, have asked:

> What if a legal stimulant that is shown to be safe could be used to improve medical care during periods of fatigue, regardless of the number of hours worked? Would not the more ethical choice be to promote the reduction of errors — First, do no harm? With publication of original research [. . .] use of prescribed drugs to promote wakefulness in the absence of comorbidities, such as narcolepsy and sleep apnoea, will undoubtedly move from the military into the public arena. (Rose and Curry 2010)

The editors think that the use of drugs to promote wakefulness is likely to move into the public arena and that this would be the ethically correct course of action in medicine, according to the professional standards guiding practitioners' conduct.

Writing in the Journal of Surgical Research, surgeons have also suggested that it may come to be required practice. They say:

> The prospect of fatigued surgeons taking a prescription drug, such as modafinil, to allow them to operate for longer, and possibly to a higher standard, is perhaps not as far-fetched as some may suggest. This drug has already been trialed in emergency physicians, when performing non-medical-related tasks at the end of a nightshift. (Warren et al. 2009)

They also emphasize that the concept of surgeons risking their health to benefit patients is not an alien one. They cite operating on patients with blood-borne transmissible diseases as an example of where the risk to the surgeon is felt justifiable to improve the patients' chances of recovery. Having noted that there are "useful and warranted forms of coercion" such as forcing surgeons to undertake hygiene practices such as handwashing prior to and during surgery, they ask:

> What will our employers feel about a drug that makes us less prone to error, able to work longer hours, or to operate more efficiently? Employers are able to request certain behavioural standards from their employees, dictate rest periods, and insist on abstinence from certain drugs to ensure that their doctors perform well — will a day arise where they can recommend or even insist on surgeons being artificially enhanced? This may seem fanciful, but recent work has suggested that a mixture of napping and caffeine attenuates fatigue in interns and thus should be adopted by hospital administration. Why not other types of stimulant?

As evidence of the plausibility of this prediction, regulators in Australia are already considering the practical dimensions of using stimulants like modafinil as a fatigue-management strategy. Queensland Health, the medical regulatory body of the North-East Australian state, in their Fatigue Risk Management System Resource Pack (2009) note:

> To meet the needs of patients at any time of the day or night, . . . doctors and other healthcare workers . . . often work long hours — throughout the night and on-call over weekends, public holidays and other times of need. This presents us with the challenge of fatigue and its associated risks to staff and patients. To meet this challenge, fatigue risk management must be included in our core business operations.

The report suggests that doctors could take "[n]aps of less than 30 min in length [t]o provide measurable boosts in alertness and performance" and up to "400 mg of caffeine [which is the] equivalent to about five to six cups of coffee" because "[c]ompared with other psychoactive drugs (e.g., modafinil), caffeine is . . . more readily available and less expensive." It should be noted that this discussion only cites the availability and expense of modafinil as counting against its use as a way to manage fatigue. Both of these barriers could be removed in the future, leaving the possibility of future reports recommending the use of psychoactive drugs like modafinil.

Further, there are domains where enhancement could already be required. Noting that the US Uniform Code of Military Justice requires soldiers to accept medical interventions that make them fit for duty, Ravelingien and Sandberg (2008) have suggested that this is at least one domain where wake enhancement may be explicitly required. They further suggest that "[i]t is conceivable that the availability of safe and effective wake enhancers will create or fortify a responsibility to ensure that fatigue no longer affects performance, particularly within professional contexts."

Whether such a responsibility will be legally enforceable is a question that has also received a tentative "yes" from legal scholarship. Chandler (2013) has argued that civil law might indirectly require surgeons to enhance, as what is viewed as delivering "reasonable care" changes. She suggests that such changes could have the effect that a failure to adopt novel neurotherapies that remedy cognitive limitations would be negligent if it could be shown to have led to harm. She emphasizes the indirect way in which this could occur:

> Cognitive deficits may also raise the risk of liability if they cause a physician to make errors that would not be made by the reasonably prudent practitioner in the field, or to fail to keep up with developments in the field to an extent that is considered to fall below the reasonable standard of care in the profession. In such cases, the courts would simply find there had been a failure to maintain the standard, without necessarily commenting on cognitive deficits or therapeutic methods to alleviate them.

A Legal Obligation to Enhance?

Could someone be legally obliged to enhance themselves? This question prompts consideration of the most likely scenario in which this might occur – a scenario in which a person has fallen below normal capacity and taking an enhancer would bring them back to normal capacity. In the scenario, the person does not take the enhancer and someone is harmed. The injured party brings a claim in negligence. This then raises three questions: Did the person have a duty to take the enhancer

(the duty question)? Is the failure to take it a breach of that duty (the breach question)? Could the failure be said to have caused the injury (the causation question)? These questions are likely to be the same questions asked in other jurisdictions, but the answers may differ. The analysis that follows presents answers based on English law.

The first question potentially involves what is called liability for a pure omission. The English common law makes a fundamental distinction between acts and omissions. Generally, if I foresee, or can reasonably foresee, that my action might harm another, I am under a duty not to act. However, the same is not true of omissions. Merely foreseeing that a *failure to act* might lead to harm does not, without more, legally *oblige* me to act. As Lord Diplock stated in *Dorset Yacht Co Ltd v Home Office* (1970):

> The very parable of the good Samaritan (Luke 10, v. 30) which was evoked by Lord Atkin in *Donoghue v Stevenson*. . . illustrates, in the conduct of the priest and of the Levite who passed by on the other side, an omission which was likely to have as its reasonable and probable consequence damage to the health of the victim of the thieves, but for which the priest and Levite would have incurred no civil liability in English law. (*Home Office v Dorset Yacht Co* 1970)

Even doctors, nurses, and other medical practitioners who come across someone in need of medical attention for whom they have not previously assumed responsibility will not be liable for failing to act (Deakin et al. 2008; *Stovin v Wise* 1996; *X v Bedforshire CC* 1995).

This reluctance stems from a commitment to protecting and promoting individual liberty, as well as equity concerns (the "why me?" objection to imposing liability on one person who does not act, while leaving another bystander not open to liability). Liability for pure omissions will be imposed in certain situations, such as where a person has created a source of danger or failed to warn about a danger where she ought to do so (see *Kane v New Forest DC* 2002) and also those in which a person is in relationship of control with a third party, where the latter causes the harm (*Home Office v Dorset Yacht Co* 1970). In addition, a person can become obliged to act if he voluntarily assumes responsibility for someone else, whether by agreeing or indicating that he will do so (*Hedley Byrne v Heller* 1964; *Barrett v MOD* 1995). It is on this basis that surgeons come under a duty to undertake positive acts in the care of their patients and so can be held liable for negligent omissions. This assumption occurs at the moment they agree to treat the patient, and from this point onwards, the surgeon must act in accordance with the accepted practices of a responsible body of medical practitioners (*Bolam v Friern Hospital* 1957). The special relationship that arises places on surgeons the duty to *do* (and refrain from doing) particular things to protect their patients' interests (Jackson 2010). This encompasses a duty not to exacerbate the situation, but the defendant will generally not be liable if the claimant is left no worse off than he would otherwise have been (*East Suffolk Rivers Catchment Board v Kent* 1941). Therefore, in such situations, it is well established that the practitioner may be liable for failing to act, and so this is one context in which liability for failure to enhance is most likely to be found, if it can be found at all. The following scenario is drawn upon to explore liability in this context.

The Scenario

A fatigued surgeon has worked 36 h without rest and is on her way home when an emergency case arrives. She is the only available surgeon qualified to perform the surgery. Despite knowing she is exhausted, she believes she can remain sufficiently capable for the duration of the surgery the patient needs. She expects the procedure to take around 2 h but recognizes that there is a risk that she might make an error due to her fatigued state. She also knows that the only other option is for the patient to be sent to the nearest hospital, which is more than an hour away. She knows that within the hour the patient's condition will deteriorate considerably and that there is the possibility that he will suffer harm as a result. She is faced with an unenviable choice but decides to perform the surgery.

As she is about to begin, she remembers that she has recently been taking a new kind of medication: modafinil. She has found that it makes her more alert when tired and enables her to concentrate better on complex tasks. Now a third choice presents itself. She could ask a nurse to get the medication for her and take it to counteract her fatigue.

The Duty Question

In our scenario, it is well established that the surgeon has a duty of care; for, as noted earlier, medical practitioners will be held liable for negligent omissions made in the course of patient care. The key question is whether a failure to take an enhancer will breach that duty.

While it is unlikely that the surgeon's employing hospital has a duty to ensure that a patient receives careful treatment,[1] it will have a primary duty to provide competent staff and proper facilities. This duty probably extends to providing properly skilled medical staff and an adequately equipped hospital, and this might include avoiding situations in which there are insufficient staff, leading to a surgeon making a fatigue-related error.[2] The hospital will also be vicariously liable for tortious omissions made by the surgeon during the course of employment. Therefore, if the surgeon is liable for failing to enhance, then the hospital that employs her will be vicariously liable on her behalf.

The Breach Question

The breach question (or put another way, the standard of care required of the surgeon) is approached in a specific manner in the context of medical treatment.

[1]See *A (A Child) v Ministry of Defence* (2004) EWCA Civ 641.

[2]See *Bull v Devon AHA* (1993) 4 Med LR 117; *Wilsher v Essex Area Health Authority*; *Re R (a minor) (No. 2)* (1997) 33 BMLR 178.

It still takes the essential risk-benefit approach, but in a way tailored to medical practice. If we ask "what must a surgeon do to fulfill her duty of care towards a patient for whom she has assumed responsibility?," the answer is provided by what has become known as the "*Bolam* test," after *Bolam v Friern Hospital* (1957). According to the test, a defendant "is not guilty of negligence if he has acted in accordance with a practice accepted as proper by a responsible body of medical [persons] skilled in that particular art." In our scenario, the question for the court would be whether a responsible medical practitioner would use enhancers in this situation. The answer at the present moment would obviously be that it is not: the use of cognitive enhancers is not mainstream practice, and a body of medical persons who would testify to this could easily be found. Therefore, this omission is unlikely to be a breach.

Since 1998, the assessment of the standard of care is no longer determined solely by a body of medical practitioners, and the court must now be ". . .satisfied that, in forming their views, the experts have directed their minds to the questions of *comparative risks and benefits* and have reached a *defensible conclusion* on the matter" (*Bolitho v City and Hackney HA* 1998). This is known as the "*Bolitho* gloss" on the test. Following *Bolitho*, the defendant's actions will not be excused merely on the basis of the support of her peers if they lacked a logical or reasonable basis. Therefore, even if the defendant could point to a body of medics who asserted that they would not use cognitive enhancers, if it became apparent that cognitive enhancers cheaply and effectively reduced risk, a judge would be able to (but would not necessarily have to) find the expert's testimony indefensible. The consequence of this is that if taking an enhancer was clearly safe and efficacious, posing little or no risk to the surgeon and having demonstrable benefits for the patient, then it is very likely that a failure to enhance would breach her duty and conversely that taking it would be the responsible course of action. Taking the modafinil would be equivalent to many other simple, non-harming precautions that a responsible surgeon would, following *Bolam* and *Bolitho*, be expected to take. Handwashing is one simple example. Wearing glasses would be another (*Stefanyshyn v Rubin* 1996).

However, at present, this could not be said for cognitive enhancers. For example, it is not yet established that modafinil is safe (Kumar 2008; McBeth et al. 2009). Most studies analyzing the effects of modafinil were small and generally did not use a standardized method for assessing adverse reactions, but there are reports of a range of side effects associated with modafinil. These include headaches, dizziness, abdominal pain, dry mouth, nervousness, restlessness, sleep disturbance, and heart palpitations. It is important to note also that the majority of studies conducted on modafinil were short-term or single-dose studies and hence do not provide much data about long-term implications and the potential for dependency (Kim 2012; Repantis et al. 2010; cf. Randall et al. 2004; Sugden et al. 2012).

Given these potential risks, and English law's resistance to requiring people to risk their own safety for others, it is unlikely there would be any expectation on a medical practitioner to take an unproven pharmaceutical. This is particularly likely given the English legal system's difficult experiences dealing with the legacy of asbestos, thalidomide, diethylstilbestrol (DES), and nicotine, which were all

considered safe until their deleterious health effects emerged, often with tragic consequences (*Fairchild v Glenhaven* 2002; Dunea and Last 2001).[3] Therefore, enhancement would be considered merely going above and beyond the call of duty, but not expected, and failing to do so would not be a breach.

These concerns are further strengthened by other, less obvious, implications of taking modafinil and other enhancers. While the drug does restore wakefulness, it does not remove the need to sleep. Users must still make up the sleep later; otherwise a "sleep-debt" is built up, which has well-established negative impacts on health. At present, the move is towards reducing doctors' hours. It is unlikely that this alternative means of robbing them of their sleep will find support, as it would go against current policy trends and also promote the normalization of a drug that leads eventually to chronic fatigue.

Additionally, once taken, one cannot simply "switch off" the effects of an enhancer. For example, modafinil loses its pharmacologic activity over time, decreasing by half around every 12–15 h (Robertson and Hellriegel 2003). Put another way, it takes a while to wear off. Consequently, a surgeon taking the enhancer will potentially remain awake for many hours beyond the time when she has completed the surgery. It will be difficult for her to fall asleep, and she will push back the time at which she goes to bed. If she has to work at her normal time the next day, she will have less sleep and be less refreshed (Gill et al. 2006). Unless she is fortunate enough to have an opportunity to catch up her lost sleep, she will very likely have to come to work at her normal time on her next working day and so will both accrue a sleep-debt and be fatigued at work.

Another concern that speaks against failure to take an enhancer being a breach is the lack of demonstrated consistent efficacy. It cannot be said that a surgeon *ought* to take the modafinil if we cannot say with sufficient certainty that doing so would be effective, as we would then be finding the surgeon in breach of duty for something that may have been irrelevant in the circumstances. The efficacy of most enhancers is yet to be conclusively proven (Lynch et al. 2011; Turner et al. 2003). Even modafinil, which appears to be the most effective, produces varied responses in different studies. Some studies do show improvements on some aspects of performance that are otherwise affected by fatigue – for example, modafinil has been shown to improve some aspects of decision making. However, there is little proof that modafinil improves the deficits in psychomotor performance that result from fatigue (Sugden et al. 2012). Similarly, while some studies showed improvements in attention, others failed to do so (Randall et al. 2004).

The effects of enhancers also vary between people in the extent to which they improve various cognitive capacities. For some, modafinil addresses fatigue effectively or leads to clear improvements in cognitive capacity. However, in others it has little or no effect (Finke et al. 2010). It is also known that the impact of modafinil on an individual varies with the context in which he takes it (Thomas and Kwong 2006). The idea of a duty to enhance rests on the notion that in doing so,

[3]We are indebted to Professors Peter Cane and Jane Stapleton for inspiring our ideas on this point.

the chances of harm are reduced. But this cannot be the case if the efficacy of the enhancer is so variable that its effects are not highly predictable.

For all of the reasons given in this section, and the general resistance to a duty of rescue, English courts are unlikely to extend a surgeon's duty to risk her own health for the sake of a patient by taking a relatively untested pharmaceutical. However, those who do take it are probably not in breach unless a clear causal link between taking the modafinil and the resulting harm can be established. The issues discussed in the next section demonstrate how difficult this would be.

The Causation Question

If the seemingly insurmountable hurdle of the second question could be overcome, the third question – causation – provides yet another barrier to finding that someone should take an enhancer or face legal liability. Under English law, a defendant cannot be liable in negligence unless it can be said that the negligent act or omission *caused* the loss that is the basis of the claim. The court must ask whether the loss would have occurred "but for," or without, the act or omission that constituted the breach (*Barnett v Chelsea & Kensington Hospital* 1968).

Applying this to the fatigued surgeon who omits to take modafinil, the question for the court would be whether "but for" the surgeon's failure to take the enhancer, the patient would have suffered the injury. In the scenario, the breach of duty was the surgeon's failure to take a cognitive enhancer. However, given that it is being assumed (plausibly) that cognitive enhancers are effective at remedying fatigue, the causal uncertainty before the court lies in the effects of persistent fatigue on surgical performance. The cause of the harm in the scenario is the fatigue, the breach is the failure to correct it, so the court would have to find that the fatigue that the surgeon failed to remedy was causally linked to the harm that eventuated.

There are two principal sources of uncertainty that might prevent the court from proving the causal link. The causal link must be shown on the balance of probabilities; that is, it must be more likely than not. The first source of uncertainty is that the effects of fatigue on performance are not clear-cut and the second is that some risks inherent in surgery can materialize blamelessly. Each of these is explained in turn.

The effect of a surgeon's fatigue (and, hence, her failure to remedy it) on surgical performance will depend on the particular procedure in question and, probably, the particular surgeon. Was it a procedure that needed expert manual dexterity? Was it a procedure that required paying attention to many things at once? Was it a procedure that required a high level of teamwork? Fatigue will affect a surgeon's ability to meet the cognitive demands of surgery to a greater and lesser extent, depending on those required for the particular procedure. Further, enhancement drugs (while having the general effect of increasing subjective wakefulness) may improve performance in some ways but have no effect on, or even impair, other domains of cognition (Repantis et al. 2010). Further still, while surgeons will have developed their own strategies for managing their fatigue, the court could not be

sure that taking an enhancer will not cause some to overestimate their capacities, perhaps *increasing* the likelihood of error (Buguet 2003). This means that the link between the uncorrected fatigue and the harm will always be speculative. Whereas, for example, not giving a patient enough oxygen or administering an overdose of a drug has known and measurable effects, the effects of operating while fatigued are indeterminate. Whereas a court can relatively confidently know what would have happened had the patient not been given an overdose, it will be very uncertain what would have happened had the surgeon taken an enhancer, i.e., whether the harm would have occurred *but for* the breach.

The second source of causal uncertainty stems from the fact there are always risks involved in surgery, which are often blameless if they materialize. These sorts of risks are explained to patients when they consent to an operation or other procedure. It would be very difficult to know whether or not some instances of harm were risks that materialized blamelessly or were consequences of the fatigue. Much will depend on the precise details of the case: the particular procedure that took place, the level of cognitive and manual dexterity involved, the risks inherent in the procedure, the probability that these risks materialize, and so on. Any risk created by the surgeon's fatigue does not remove or subsume the risks that were already present. The surgeon will not be liable for these risks if she has adequately explained them to the patient prior to obtaining consent to the surgery (but will remain potentially liable for any risks that arise due to her failure to take due care). Consequently, there are some risks that can occur blamelessly (i.e., not due to negligence, but simply as part of the normal procedure) and for these the surgeon will not be blamed (if fully informed consent has been obtained). Given this, it would be unfair to assume a link to fatigue where the surgeon omits to enhance. If some harms can occur blamelessly when the surgeon is alert, why not also when fatigued?

So, establishing causation would require the court to clearly conceive the causal mechanism from breach to injury, understanding the role which uncorrected fatigue might play in it. Interpersonal differences in the experience of fatigue and the uncertainty about its precise effects on surgical performance make this a huge empirical challenge. The assessment is further confounded by the fact that the risks inherent in surgery often materialize blamelessly. Even on the balance of probabilities standard, courts would likely be unable to establish causation.

When Enhancers Cause Harm

Our focus has been on the failure to enhance up to normal capacity, but it is possible that a claim might be brought for harm caused by the decision to enhance. This would operate in the same way whether it was to bring the person up to their normal level of capacity or beyond it. The first question is whether the defendant was under a duty to take care in the situation. To answer this generally, or in the abstract, is difficult. Duties of care relate to the context in which the act is done, and there is often existing case law that determines whether an act can be considered negligent

in context. Therefore, the question cannot be answered in the abstract in any meaningful way.

The vital question, however, will be the second question: could taking an enhancer be said to breach the duty? Put another way, would taking an enhancer fall below the expected standard of care? Would it be something a responsible surgeon would not do? Necessarily, this standard is related to the duty imposed in the context, but some general comments can be made about how the question might be approached. Breach is determined by assessing the balance of risks in a given situation. In considering how to act, one must weigh the magnitude of the foreseeable risk that the harm will eventuate if one acts. Where the risk of harm is high, and the cost of avoiding this risk is low, carrying out the risky action will be a breach of duty. Conversely, where there is a very low risk of harm, and the cost of not acting is high, then it may be reasonable to act and there will be no breach (*Miller v Jackson* 1977; *Bolton v Stone* 1951).

How would this calculus be applied where a person takes a cognitive enhancer? It would depend on the context, and it is difficult to analyze in a vacuum, so the focus will be on the surgeon again. Supposing she chooses to take a cognitive enhancer while at work because it helps her to concentrate, and it will be presumed that the enhancer will definitely have this effect. For this action to be injurious to another (and hence open the door to her being liable in negligence), taking the enhancer must allegedly be harmful to someone else. This will be very difficult, as taking the enhancer only directly affects herself. Perhaps it could be imagined that in becoming so focused, she tends to focus on one task to the exclusion of others. This leads to her to ignore some aspects of patient care, and the patient is harmed as a result. Alternatively, she takes the enhancer to ensure she can stay awake during a very long surgery, but this means she becomes too focused on one aspect of the surgery and neglects another, leading to a harm. In either case, the law would weigh the risks to the patient (neglecting patient care) against the costs of not taking it (i.e., the benefits of taking it such as improved concentration). The best that might be said in such an abstract situation is that the risk of harm, if very serious, would quite likely outweigh the benefits given that most doctors manage to care successfully for patients without enhancing themselves.

The third question, the causation question, also poses significant difficulties. As explained earlier, to prove causation, it must be shown that but for the breach of duty, the harm would not have occurred on the balance of probabilities. Where the breach is the taking of an enhancer, a causal link between taking that enhancer and the resulting harm needs to be shown. We must be able to say that but for someone taking the enhancer, the harm to the claimant would not have occurred. The enhancer would therefore need to have some harming effect in this context. If modafinil is taken as the example, as seen above, it could be speculated that it made the surgeon too focused, but a surgeon is already concentrating very hard in the context of surgery. It is self-evidently almost impossible to show that it was, more likely than not, only the *that extra* concentration attributable to the modafinil (as this is the breaching act) that led her to ignore an aspect of patient care, which in turn led to the harm complained of.

The other possible harming effect of taking modafinil that might be imagined is the accrual of a sleep-debt noted above. It could be argued that taking the modafinil earlier caused the surgeon to experience the effects of the sleep-debt during this patient's surgery and thus become fatigued and make an error related to that fatigue. This is again highly likely to fail, partly because it falls foul of the problems noted in the previous section on linking fatigue to harm but compounded by the many intervening factors between the original taking of the modafinil and the later sleep-debt and consequent error. Many breaks in this causal chain could be suggested, and establishing causation would be difficult, if not impossible. Even if factual causation could be established, the remoteness limitation that restricts liability to reasonably foreseeable consequences would also very likely prevent liability being found in such a situation (*Overseas Tankship v Morts Dock Ltd* 1961).

Does Enhanced Capacity Give Rise to Enhanced Legal Responsibility?

A related question is whether surgeons who were cognitively enhanced should be held more responsible – held to higher standards. The assumption underlying this idea is that as a person's mental capacities increase, so does their level of responsibility. Nicole Vincent has articulated this "capacitarian" thesis and illustrates it as follows:

> [I]n lay contexts responsibility is often thought to require such things as the ability to perceive the world without delusion, to think clearly and rationally, to guide our actions by the light of our judgments, and to resist acting on mere impulse. This is, for instance, why children, the senile, and the mentally ill are thought to be less than fully responsible for what they do (i.e. because they lack the right kind and/or degree of mental capacity), why children can acquire more and/or greater responsibilities as they grow up (i.e. because their mental capacities develop as they mature), and how responsibility is reinstated on recovery from mental illness (i.e. because the needed mental capacities are recovered). (Vincent 2011)

Elsewhere (Vincent 2013), she suggests that the implication of the capacitarian thesis is that as capacities are enhanced beyond the "normal" range, the people in possession of these capacities might in some sense become "hyper-responsible." This means that these people should be held *more* responsible – are more blameworthy – when things go wrong. If this theory is correct, then there are obvious implications for surgeons who use enhancers: once enhanced, a higher level of proficiency can be expected of them. Would enhanced surgeons become liable for errors that previously would have been tolerated as results of inescapable human error? While this remains an interesting question for moral philosophers, the law is unlikely to reflect the capacitarian implications. In negligence law, an individual's precise capacity is not generally relevant to the existence of a duty and its breach, as the standard is the objective standard of a reasonable person. The learner driver is held to the same standard as the professional racing driver. In the context of surgeons, the standard of care will be that of the reasonable surgeon. It is unlikely that a different standard – that of the reasonable enhanced surgeon – would be

expected from those surgeons that enhanced. However, were enhancement to become mainstream practice, and were there to be a consistently higher level of proficiency as a result, it is conceivable that what is expected of the reasonable surgeon might rise. This happens with the introduction of all different kinds of technology and practices to clinical care: the standards expected now are higher than those expected 50 years ago, but this happens across the board and is not idiosyncratic to individual surgeons' cognitive or other capacities.

At present, English negligence law remains tied to responsibility as a threshold concept, determined by either reasonableness or the responsible practitioner standard, and there is little to suggest that this will change in the near future. However, whether the capacitarian approach will find a foothold remains to be seen, and there is much further useful work to be undertaken on the questions raised in this section.

The Right Role of the Law: Legal Responsibility, Ethical Obligations, and the Demands of the Public

The arguments put in the ethical and philosophical literature in favor of professionals being obliged to enhance themselves may well be sound. It might well be the case that it would be the right thing to do, ethically speaking, for a tired surgeon to take modafinil to enable her to take on a surgery she might otherwise not be able to perform. A pilot might be morally obliged to enhance himself to reduce the risk of errors, given the potentially catastrophic consequences if he cannot safely pilot the plane. So too the public might legitimately wish these professionals to self-enhance to help prevent the possibility of harms. Very often, the law will follow or be strongly influenced by such ethical obligations and the views of the public. However, for two main reasons, this will not always be the case.

The first of these rests on the role of the law in society. The law is an organizing system, guiding our behavior to enable us to achieve our individual and collective goals. Although it depends on one's jurisprudential stance, it is probably uncontroversial to hold that the law should afford us the greatest amount of liberty that can be achieved while still protecting the liberty and interests of other members of society. One might go further and argue, as Joseph Raz does, that the law's role is not simply to facilitate liberty and pure choice but to promote autonomy in the rich sense of choices that contribute to human flourishing. On this view, the law ought to offer us a range of *good* choices, which should be more attractive than those that are less likely to promote our interests (Raz 1986). Whatever stance is taken, it is also uncontroversial that the law should be sufficiently clear and stable to effectively guide behavior – the rule of law demands that citizens are able to determine with reasonable certainty what they may and may not do.

To find liability in the context of enhancement, then, the law would need to be able to very clearly define when enhancement would be required. This requires a degree of certainty that the ethical obligations described earlier in this chapter lack. In situations where the risk to the surgeon from enhancement is low, and the risk to the patient otherwise very high, it will be clear that there is a moral

obligation to enhance and save the patient. However, there will be many situations in which this is not so. While it might be obvious that a passerby is morally obliged to turn over the drowning baby in the puddle because there is no risk to himself, the determination of risk is exponentially more difficult when the puddle becomes an ocean. It cannot be said with certainty that the risk in saving the baby in the ocean would be negligible and so would arrive at a situation where the law would be compelling one person to risk themselves for another. This would require not only a risk-benefit analysis as discussed above but would further demand that the law could determine the precise level of risk that a bystander ought to take on to help another person (in the absence of any assumed responsibility). Such precision would be crucial, because it would determine whether or not the bystander would be liable for the resulting harm (from not aiding the person in peril) and would also deem that bystander's moral failure *the legal cause* of the harm, despite his lack of involvement in the situation in any other causal sense. These points, coupled with the English legal system's commitment to promoting and protecting individual autonomy, leads it to eschew requiring people to act even though it would be moral to do so. This is strengthened by the fact that a failure to do as the law demands in such a case does not lead just to moral disapprobation but sanctions. These might be a requirement to compensate but in some contexts might involve punitive measures and the labeling of the person who did not risk themselves for another as a criminal.

This divergence between ethics and law arises further because the law must apply to everyone and it must be discoverable and sufficiently certain as to be an adequate guide to behavior. It cannot so easily take account of the nuances of a particular situation like that in the preceding paragraph, so it must draw a line and put forward a principle that can be definitely stated and stood by. For this reason, the law will not always map directly onto the terrain so carefully teased out by the ethics of a situation. In the context of medicine, the contrast in approaches is spelt out neatly by a comparison of the Hippocratic Oath – first, do no harm – and the law's position, do not make the situation worse. The question of what constitutes harm in any given situation will be complex, subtle, and potentially vexed. A directive not to make things worse (or otherwise face sanction) is more certain, perhaps more blunt, but probably a more workable guide to behavior that would otherwise attract legal response.

The law must also demand or prohibit behavior to ensure the protection of all citizens. If self-enhancement carries risks, then this must come into the equation not only in the individual case as above but also when considering the wider, systemic effects of expecting individuals to enhance themselves. As seen above, if the law finds that failing to enhance is a breach of duty, this necessarily creates a widespread obligation on many professionals to regularly self-enhance to avoid the possibility of being found legally liable for harms. This shifts the norms in their workplaces, and a shift to regular enhancement as the workplace norm would consequently place significant strain on individuals. This might also lead to defensive practices where, for example, doctors' choice of patient treatment is determined in part by a desire to avoid liability, rather than solely based on what is most likely to cure.

Conclusion

This chapter has briefly explored one context in which calls have been made for a particular group of professionals to begin using new enhancement technologies for the benefit of others. While there is much, morally, to recommend surgeons doing their best to save others where the cost to themselves is low, the law cannot easily take this position. For good reasons, the English common law resists imposing liability (and consequent sanctions and demands for compensation) on those who fail to act. For surgeons, they are to be judged by a standard determined by their peersand by the evidence in support of a practice. At present, such a standard would not, in our view, demand that a surgeon take an enhancing drug. The law also demands a high degree of probability that an action has caused the harm complained of. Such an approach is only fair when the defendant faces paying significant compensation if the causal connection is made. The law must be sure that responsibility has been placed on the right shoulders. Given the state of the science, and our capacity to determine why a harm has arisen, it is highly unlikely that such a causal connection could be established between either taking an enhancer when the surgeon should not or a failure to take when she should. It is, for the reasons given in this chapter, even more unlikely that a surgeon will be expected to enhance herself beyond the norm. Given this, it should be concluded that, at the current time, English law is unlikely to find a surgeon who takes, or fails to take, an enhancer liable in negligence.

Cross-References

▶ Cognitive Liberty or the International Human Right to Freedom of Thought
▶ Neuroenhancement
▶ Neurolaw: Introduction
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play
▶ Reflections on Neuroenhancement
▶ Smart Drugs: Ethical Issues
▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Articles and Books

Academy of Medical Sciences, Royal Society, British Academy, Royal Academy of Engineering. (2012). Human enhancement and the future of work (Report from Joint Workshop). http://www.acmedsci.ac.uk/p47prid102.html#downloads. Accessed 22 May 2013.

Baranski, J. V., Gill, V., McLellan, T. M., Moroz, D., Buguet, A., & Radomski, M. (2002). Effects of modafinil on cognitive performance during 40 hr of sleep deprivation in a warm environment. *Military Psychology, 14*, 23–47.

Buguet, A. (2003). Modafinil – Medical considerations for use in sustained operations. *Aviation, Space, and Environmental Medicine, 74*, 659–663.

Caldwell, J. A. (2005). Dextroamphetamine and modafinil are effective countermeasures for fatigue in the operational environment. US Air Force Research Laboratory. http://ftp.rta.nato.int/Public/PubFullText/RTO/MP/RTO-MP-HFM-124/MP-HFM-124-31.pdf. Accessed 22 May 2013.

Chandler, J. A. (2013). Autonomy and the unintended legal consequences of emerging neurotherapies. *Neuroethics, 6*, 249–263.

Cooper, M. R., Bird, H. M., & Steinberg, M. (2009). Efficacy and safety of modafinil in the treatment of cancer-related fatigue. *Annals of Pharmacotherapy, 43*, 721–725.

Deakin, S., Johnston, A., & Markesinis, B. (2008). *Markesinis and Deakin's tort law*. Oxford: Oxford University Press.

Dunea, G., & Last, J. M. (2001). *The Oxford illustrated companion to medicine*. Oxford: Oxford University Press.

Finke, K., Dodds, C. M., Bublak, P., Regenthal, R., Baumann, F., Manly, T., & Müller, U. (2010). Effects of modafinil and methylphenidate on visual attention capacity: A TVA-based study. *Psychopharmacology, 210*, 317–329.

Gill, M., Haerich, P., Westcott, K., Godenick, K., & Tucker, J. (2006). Cognitive performance following modafinil versus placebo in sleep-deprived emergency physicians: A double blind randomized crossover study. *Academic Emergency Medicine, 13*, 158–165.

Grady, S., Aeschbach, D., Wright Jnr, K. P., & Czeisler, C. A. (2010). Effect of modafinil on impairments in neurobehavioral performance and learning associated with extended wakefulness and circadian misalignment. *Neuropsychopharmacology, 35*, 1910–1920.

Hart, C. L., Haney, M., Vosburg, S. K., Comer, S. D., Gunderson, E., & Foltin, R. W. (2006). Modafinil attenuates disruptions in cognitive performance during simulated night-shift work. *Neuropsychopharmacology, 31*, 1526–1536.

Jackson, E. (2010). *Medical law: Text, cases and materials*. Oxford: Oxford University Press.

Kim, D. (2012). Practical use and risk of modafinil, a novel waking drug. *Environmental Health and Toxicology, 27*, e2012007 (ePub pages only).

Kumar, R. (2008). Approved and investigational uses of modafinil: An evidence-based review. *Drugs, 68*, 1803–1839.

Lynch, G., Palmer, L. C., & Gall, C. M. (2011). The likelihood of cognitive enhancement. *Pharmacology, Biochemistry, and Behavior, 99*(2), 116–129.

McBeth, B. D., McNamara, R. M., Ankel, F. K., Mason, E. J., Ling, L. J., Flottemesch, T. J., & Asplin, B. R. (2009). Modafinil and zolpidem use by emergency medicine residents. *Academic Emergency Medicine, 16*, 1311–1317.

Randall, D. C., Fleck, N. L., Shneerson, J. M., & File, S. E. (2004). The cognitive-enhancing properties of modafinil are limited in non-sleep-deprived middle-aged volunteers. *Pharmacology Biochemistry and Behavior, 77*, 547–555.

Ravelingien, A., & Sandberg, A. (2008). Sleep better than medicine? Ethical issues related to "wake enhancement". *Journal of Medical Ethics, 34*, e9–e9.

Raz, J. (1986). *The morality of freedom*. Oxford: Oxford University Press.

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*, 187–206.

Robertson, P., & Hellriegel, E. T. (2003). Clinical pharmacokinetic profile of modafinil. *Clinical Pharmacokinetics, 42*, 123–137.

Rose, S., & Curry, T. (2010). Fatigue countermeasures, and performance enhancement in resident physicians – Reply. *Mayo Clinic Proceedings, 85*, 301–302.

Roth, T., Schwartz, J. R. L., Hirshkowitz, M., Erman, M. K., Dayno, J. M., & Arora, S. (2007). Evaluation of the safety of modafinil for treatment of excessive sleepiness. *Journal of Clinical Sleep Medicine, 3*, 595–602.

Sandberg, A., Sinnott-Armstrong, W., & Savulescu, J. (2011). Cognitive enhancement in courts. In J. Illes & B. J. Sahakian (Eds.), *Oxford handbook of neuroethics* (pp. 273–284). Oxford: Oxford University Press.

Sugden, C., Housden, C. R., Aggarwal, R., & Sahakian, B. (2012). Effect of pharmacological enhancement on the cognitive and clinical psychomotor performance of sleep-deprived doctors: A randomized controlled trial. *Annals of Surgery, 255*, 222–227.

Thomas, R. J., & Kwong, K. (2006). Modafinil activates cortical and subcortical sites in the sleep-deprived state. *Sleep, 29*, 1471–1481.

Turner, D. C., Robbins, T. W., Clark, L., Aron, A. R., & Dowson, J. (2003). Cognitive enhancing effects of modafinil in healthy volunteers. *Psychopharmacology, 165*, 260–269.

Vincent, N. A. (2011). Capacitarianism, responsibility and restored mental capacities. In *Proceedings of the conference Technologies on the stand: Legal and ethical questions in neuroscience and robotics*, Tilburg University, pp 41–62.

Vincent, N. (2013). Enhancing responsibility. In N. A. Vincent (Ed.), *Neuroscience and legal responsibility* (pp. 305–334). New York: Oxford University Press.

Warren, O. J., Leff, D. R., Athanasiou, T., Kennard, C., & Darzi, A. (2009). The neurocognitive enhancement of surgeons: An ethical perspective. *Journal of Surgical Research, 152*, 167–172.

Cases

A (A Child) v Ministry of Defence (2004) EWCA Civ 641.

Barnett v Chelsea & Kensington Hospital Management Committee (1968) 1 All ER 1068.

Barrett v MOD (1995) 1 WLR 1217.

Bolam v Friern Hospital Management Committee (1957) 1 WLR 582.

Bolitho v City and Hackney Health Authority (1998) AC 232.

Bolton v Stone (1951) AC 850.

Bull v Devon AHA (1993) 4 Med LR 117.

East Suffolk Rivers Catchment Board v Kent (1941) AC 74.

Fairchild v Glenhaven Funeral Services (2002) UKHL 22, (2002) 3 WLR 89.

Hedley Byrne v Heller & Partners (1964) AC 465.

Home Office v Dorset Yacht Co (1970) AC 1004.

Kane v New Forest DC (2002) 1 WLR 312.

Miller v Jackson (1977) 3 WLR 20.

Overseas Tankship (U.K.) Ltd v Morts Dock & Engineering Company Ltd (1961) UKPC 1, (1961) 2 WLR 156.

Re R (a minor) (No. 2) (1997) 33 BMLR 178.

Stovin v *Wise* (1996) 3 WLR 389.

Wilsher v Essex Area Health Authority (1988) AC 1074.

X v Bedforshire County Council (1995) 2 AC 633.

The Use of Brain Interventions in Offender Rehabilitation Programs: Should It Be Mandatory, Voluntary, or Prohibited?

86

Elizabeth Shaw

Contents

E. Shaw
School of Law, University of Aberdeen, Aberdeen, UK
e-mail: eshaw@abdn.ac.uk; E.Shaw-2@sms.ed.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_169,

Abstract

As our understanding of the brain increases, it seems likely that new biomedical techniques for altering human behavior will be developed. One potential application of such techniques is within the context of the rehabilitation of offenders. This prospect may seem attractive, given the social and economic costs of crime and the limited effectiveness of traditional punishments at reducing reoffending (see, e.g., Ministry of Justice. *Proven re-offending statistics, quarterly bulletin*. October 2010–September 2011. England and Wales. Available at: https://www.gov.uk/government/publications/proven-re-offending–2, 2013). On the other hand, the fact that the brain is so intimately connected with an individual's identity, personality, and capacity for making autonomous choices gives rise to distinct concerns about the use of directly interfering with brain functioning. This chapter begins by indicating a number of neuroenhancements that might be used in rehabilitation programs in the future, if sufficiently safe and effective. Secondly, it outlines and replies to some of the main arguments for prohibiting neuroenhancement of offenders altogether. Thirdly, it examines the opposite position – that neuroenhancement of offenders should be mandatory. Fourthly, it argues that the case for mandatory interventions is ultimately unpersuasive. Provided sufficiently safe and effective interventions are developed, the best policy would be to provide such interventions on a voluntary basis. The chapter ends by discussing some of the issues that need to be addressed to ensure that offenders' consent to neurointerventions is truly voluntary.

Introduction

Currently, pharmacological and surgical means of affecting brain function are primarily used to treat mental and neurological disorders. However, as neuroscience progresses, we may increasingly be able to use brain interventions to alter behaviors that are "medically unremarkable but socially undesirable" (Farah 2004). For instance, brain interventions may be employed as part of offender rehabilitation programs and potentially offered as a condition of early release from prison. This prospect may seem attractive, given the social and economic costs of crime and the limited effectiveness of traditional punishments at reducing reoffending (see, e.g., Ministry of Justice 2013). On the other hand, the fact that the brain is so intimately connected with an individual's identity, personality, and capacity for making autonomous choices gives rise to distinct concerns about directly interfering with brain functioning.

The aim of using brain interventions in rehabilitation programs is largely to benefit society, by improving the moral conduct of offenders, rather than to further the best *medical* interests of the recipient. (For that reason, this chapter will employ the terms "neuroenhancement" or "neurointervention" rather than "treatment.") This also raises important issues about the appropriate relationship between the state and the individual and the circumstances under which the individual's interests could be overridden by the needs of wider society.

This chapter will begin by indicating a number of neuroenhancements that might be used in rehabilitation programs in the future, if sufficiently safe and effective. Secondly, it will outline and reply to some of the main arguments for prohibiting neuroenhancement of offenders altogether. Thirdly, it will examine the opposite position – that neuroenhancement of offenders should be mandatory. Fourthly, it will argue that the case for mandatory interventions is ultimately unpersuasive. Provided sufficiently safe and effective interventions are developed, the best policy would be to provide such interventions on a voluntary basis. This chapter will end by discussing some of the issues that need to be addressed to ensure that offenders' consent to neurointerventions is truly voluntary.

Examples of Potential Neuroenhancements

Increasing Empathy

There is some evidence to suggest that the ability to empathize is key to understanding moral norms (Blair et al. 2005). For example, individuals with markedly reduced levels of empathy have exhibited difficulties in distinguishing conventional rules (such as rules of etiquette) from moral rules and in ranking wrongs in order of seriousness. Philosophers differ as to whether empathy is essential for moral understanding. However, even those who believe that it is not essential often maintain that it is indirectly helpful in moral development. If techniques were produced which increased empathy in individuals who appear to be deficient in it, then this might play a useful role in reforming offenders.

Decreasing Violent Urges

Certain offenders may experience repetitive violent fantasies and powerful surges of anger which they find difficult to control. These factors can impair offenders' ability to think clearly about how they should act and may distort their moral judgments. Research is beginning to uncover certain neurological factors that seem to have an impact on individuals' dispositions to anger and violence. There is some evidence that selective serotonin reuptake inhibitors (SSRIs) may reduce aggression (Ferari et al. 2005; Douglas 2008; Crockett et al. 2010). It may become possible in the medium-term future to develop techniques which can reduce the strength of offenders' volatile impulses or which increase their control over these impulses.

Anti-libidinal Medication

Drugs have already been developed to help reduce deviant sexual urges and thoughts. This can create an opportunity for offenders to concentrate on the reasons why they should change their behavior and the steps they need to take, without being distracted by their impulses. These medications are already being used to some extent within the criminal justice system (Harrison and Rainey 2013).

Decreasing Racist Sentiments

Certain individuals experience a strong negative emotional reaction to members of different races. Such emotional reactions may stem from early childhood experiences, e.g., parents who taught them to fear members of a different race. Such deeply rooted emotional reactions may help to fuel racially motivated crimes and may interfere with the racist's ability to see why racism is wrong. Some research has been undertaken into the neural basis for racial stereotyping (Hart et al. 2000; Phelps et al. 2000; Cunningham et al. 2004). Potentially this might lead to interventions that could attenuate such emotional responses (see, e.g., Terbeck et al. 2012 on the use of propranolol in this context). Harris has criticized this proposal on the basis that racism is likely to involve a complex network of beliefs and not merely emotional reactions (Harris 2011). Although this is almost certainly true, it does not demonstrate that ingrained emotional reactions do not contribute to the tendency to hold stubbornly onto ill-founded beliefs in the face of the evidence. Attenuating such emotional responses might help the offender to assess the issues dispassionately and realize that his racist views are ill-founded (see also Douglas 2013a).

Delaying Gratification

Difficulties with delaying gratification may lie behind some individuals' tendency to break the law. Neuroenhancements could potentially help to rehabilitate criminals through enabling them to work out and implement strategies to delay gratification (see, e.g., Penney 2012).

Increasing the Ability to Focus on Relevant Issues

Recent studies suggest that individuals who score highly on measures for psychopathy may suffer from a kind of attention-deficit disorder which may help to explain their characteristic antisocial behavior (Newman et al. 2010). It seems that when presented with incentives for performing an action, these individuals lose sight of the reasons against performing the action. Neuroenhancements might enable these individuals to focus on all the relevant considerations (and in particular, the reasons against breaking the law).

Should Neuroenhancement of Offenders Be Prohibited Altogether?

This section will consider some concerns people may have about the neuroenhancement of offenders. It will argue that these concerns do not justify prohibiting such interventions, provided that offenders consent to them and that they are sufficiently safe and effective.

Neurointerventions Could Prevent Morally Desirable Behavior

In a recent article, John Harris argues that neurointerventions designed to diminish aggression may not be morally desirable (Harris 2011). He cites an example of an individual who attacked a terrorist who was about to detonate a bomb, thereby rescuing a plane full of people. According to Harris, if the rescuer had been given SSRIs to reduce his aggression, he might not have managed to save the plane.

Harris does highlight a genuine concern – moral understanding and morally motivated behavior are complex phenomena. Even the well-intentioned use of biomedical interventions risks causing undesirable consequences. However, when assessing whether offenders should receive such interventions, it is important to take into account the likelihood of the relevant scenarios occurring. In the case of many violent offenders, the risk that the intervention will prevent them from heroically rescuing a crowd of innocent people may seem relatively small compared with the risk that without the intervention, they will reoffend.

Neurointerventions Threaten Free Will/Autonomy

Harris (2011) has also argued that neurointerventions could deprive the offender of the ability to choose to do wrong – an ability that he claims is central to free will and moral agency (see also Matthews 2007).

It should be noted that this objection could only apply to neurointerventions that make it impossible to give in to temptation. However, few interventions would have that effect. Interventions that reduced the strength of some of the offender's antisocial urges increased the strength of competing desires to obey the law, or otherwise strengthened his capacity for self-control would not fall foul of Harris's objection.

Furthermore, it is doubtful that the ability to commit certain immoral acts is central to one's status as a moral agent. For instance, someone with a proper respect for the rights of women and children and with a very vivid appreciation of the horror of sexual abuse from the victim's perspective may be psychologically incapable of committing rape or acts of pedophilia. This inability would not undermine such a person's status as a moral agent.

If the offender's inability to commit certain acts is the result of his autonomous choice to receive neurointerventions (in the knowledge they would have that effect), then his inability should not necessarily be seen as an interference with his autonomy. His behavior post-intervention can be considered autonomous since it can be "traced" to his earlier autonomous decision (cf. Timpe 2011). Some offenders actively seek neurointerventions. It seems more respectful of their autonomy to offer them this option than to deny them the choice, especially in those cases where offenders have consistently requested neurointerventions over a period of time and their desire is based on full information and coherent reasons (see, e.g., McMillan 2013).

Nevertheless, certain interventions that have *globally* undermining effects on autonomy may be genuinely problematic. J.S. Mill, for instance, argued that it is impermissible to sell oneself into slavery (Mill 1909). By analogy, the individual might lack the authority to undergo neurointerventions that prevented reoffending by depriving him of the ability to make any further rational choices (cf. Bomann-Larsen 2013; McMillan 2013). Even if successful, this argument would only rule out the most extreme types of intervention.

Resisting Temptation Without Biomedical Interventions Is Intrinsically Valuable

Some writers have argued that neurointerventions could make it "too easy" to resist temptation, even if they did not eliminate the possibility of wrongdoing (Olsen 2006; Sorensen 2010). One rationale for this position is based on the value of effort itself. The attempt to address one's darker urges by reflecting on their source (e.g., in one's childhood) rather than opting for drug treatment might also lead to valuable self-knowledge.

Even granting that effort and self-knowledge are intrinsically valuable, that does not mean that *every activity* should be effortful or must generate self-knowledge. These goods are only two among many. Given that the offender's behavior can cause himself and others great harm, then surely the value of harm prevention must be given considerable weight when deciding whether interventions to address this behavior are permissible. Furthermore, even if one is primarily concerned with effort and self-discovery, *denying* an offender neurointerventions could interfere with these values. If the offender is constantly wrestling with violent or sexually abusive urges, then he may lack the time or energy to engage in other activities (e.g., charity work, the development of a talent, or relationships) that may require moral effort and may lead to self-knowledge (Levy 2006).

Neurointerventions Could Allow Offenders to Escape Their Just Deserts

Some retributivists might oppose offering neurointerventions as a condition of early release from prison on the basis that sentences should be based solely on the offenders' just deserts. Neurointerventions may render an offender non-dangerous, but this, it might be claimed, is no reason to release him if he has not yet paid the appropriate penalty for his immoral conduct.

There is not space within this article to defend a particular justification of punishment. However, it should be noted that many theorists believe that a legitimate function of the criminal justice system is to protect the public and to rehabilitate offenders. Certainly, in practice, most criminal justice systems give some weight to such goals, and are not solely concerned with retribution. Even on certain retributive views, there might be an argument for reducing an offender's

sentence if he volunteers to undergo a neurointervention. For example, the offender's request of a neurointervention might be seen as indicating his recognition of the unacceptability of his conduct (and as an attempt to prevent it recurring). This might merit a reduced sentence for similar reasons that an early guilty plea, or expressions of remorse, might be mitigating factors. On a communicative retributive view (e.g., Duff 2001) – according to which the criminal justice system should attempt to engage with dialogue with offenders and to bring them to repent their wrongful conduct and reform – neurointerventions might be seen as playing a complementary role alongside traditional punishments (see Shaw 2011). Certain neurointerventions might play this role if they could enable offenders to engage more effectively in dialogue and to empathize with their victims.

Arguments for Mandatory Neuroenhancement of Offenders

Most theorists who have written on this topic assume that the use of neurointerventions in rehabilitation programs could only be permissible if it were voluntary, without giving the possibility of mandatory interventions much attention. However, some jurisdictions currently subject offenders to compulsory biomedical interventions that have neurological effects. For instance, in certain American states, repeat sex offenders can be given mandatory chemical castration or can be forced to choose between surgical or chemical castration, both of which (as noted above) cause significant physiological and psychological changes including reducing sexual thoughts and desires (Harrison and Rainey 2013). The case in favor of involuntary neuroenhancement of offenders therefore merits further consideration. This section will attempt to put that case at its strongest. The next section will present the arguments against involuntary neurointerventions – arguments that, on balance, seem more persuasive.

Neuroenhancements May Be More Humane than Traditional Punishment

Most theorists believe that the state is entitled to subject offenders to punishment without requiring the offender's actual, explicit consent (cf. Ryberg and Petersen 2013a). Why should rehabilitation programs be treated differently? Theorists who oppose mandatory neuroenhancement are often motivated by concern for the offender's interests. However, it is possible to imagine that types of neurointervention may be developed that would cause the offender less suffering and deprivation of liberty than traditional punishment. Prisons are dangerous and depressing places that can leave offenders with permanent psychological and physical injuries. If safe and effective neuroenhancements were developed that allowed criminals to be quickly rehabilitated and returned to their communities, then it would seem more humane to compel offenders to undergo these interventions than to spend years in prison.

A Rational Offender Would Consent to Neuroenhancement

Some theorists justify punishment on the basis that, hypothetically, the offender would consent to it, if he were rational. For example, Benjamin Vilhauer (2009) has argued that society's treatment of offenders should be governed by rules that everyone would consent to from behind John Rawls's "veil of ignorance" (Rawls 2005). The veil of ignorance is a thought experiment that involves imagining oneself choosing the rules of society in accordance with rational self-interest, but without knowing what social position one will be allotted, including (as Vilhauer insists) whether one will be a victim or an offender. Decision-makers behind the veil of ignorance must therefore put themselves in the shoes of everyone affected by a social policy. This thought experiment is meant to exclude personal biases and ensure that the interests of all members of society are given equal weight. If certain neuroenhancements turn out to be safer and more effective than traditional punishment, then compulsory neuroenhancement might be in the interests of victims and perpetrators alike. If so, then the offender himself (provided he were wholly rational and imagined himself to be behind the veil of ignorance) would agree that neuroenhancement should be mandatory.

Offenders Forfeit the Right to Refuse Neuroenhancement

Alternatively, it might be argued that we should not be too concerned about protecting the interests of offenders. Various theorists claim that offenders, by committing a criminal act, forfeit their right not to be punished (e.g., Wellman 2012). It might be further argued that offenders also forfeit their right to choose whether or not to receive neuroenhancements that would address their offending behavior.

Potential Victims Are Entitled to Effective Protection

Julian Savulescu and Ingmar Persson (2008) have argued that if safe and effective biomedical moral enhancements were developed, then they should be compulsory for everyone. They believe this would reduce the chance that humanity would abuse powerful destructive technologies (e.g., nuclear materials and instruments of bioterrorism) – technologies that a rapidly increasing number of people will have access to in the future. Vojin Rakic (2013) opposes compulsory enhancement of the whole population, but suggests that it might be appropriate for certain repeat offenders, such as child rapists, whose release from prison would be dangerous. One way of specifically addressing Savulescu's and Persson's concerns might be the mandatory neuroenhancement of those who have already breached antiterrorism laws.

The Rights of Offenders' Families

A long prison sentence may separate offenders from dependent relatives, thus depriving children, partners, and elderly parents of emotional and financial support. If safe and effective neurointerventions were developed, allowing offenders to be quickly rehabilitated, then offenders could fulfill their familial obligations.

An Economic Argument

Prison is phenomenally expensive. It costs £40,000 to keep a single in-mate in prison in the UK for only 1 year (Adebowale 2010). (This is more expensive than educating a boy at Eton College for the same period.) For many offenders, prison is inadequate deterrent. According to recent Ministry of Justice Statistics for England and Wales, 11,000 offenders, who had each been jailed at least 11 times in the past, together were responsible for 50,000 offenses within a single year (Ministry of Justice 2013). If effective neurointerventions were developed, then their mandatory use in offender populations could generate massive savings. Such interventions could also benefit the economy by allowing offenders to re-enter the workforce. It might be thought that economic reasons are insufficiently weighty to justify compulsory neurointerventions. However, it should be remembered that our society is currently prepared to impose great burdens on individuals for economic reasons, e.g., withdrawing funding for potentially life-saving treatments.

Arguments Against the Mandatory Neuroenhancement of Offenders

The Power to Impose Mandatory Neurointerventions Is Too Easily Abused

There is already a vast inequality of power between the state and individual citizens. The state has the ability to use coercive force to deprive offenders of their property, cut them off from their friends, families, and communities, and to severely constrain their movements. Rather than being an argument for increasing the state's power still further, this should lead us to question whether it is justifiable to make the gap between the state and individual citizens even wider. If the authorities were able to impose mandatory neuroenhancement, this would give them a significant level of control over the individual's inner life. This obviously has potential for abuse, e.g., it could be used to suppress legitimate dissenters. By definition, it is wholly unacceptable to impose any form of coercive sanction on such dissenters. However, at least traditional methods of punishment do not interfere with these citizens' freedom of conscience to the same degree as mandatory neurointerventions and dissenters may use their time in prison as a form of protest.

During a substantial part of the twentieth century, several states in the USA had many discriminatory criminal laws, including a ban on interracial marriage (that particular law was not overturned until Loving v Virginia 388 US 1 (1967)). This great injustice would have been exacerbated if the authorities been able to force individuals convicted under such laws to undergo neurointerventions designed to decrease their attraction to members of a different race. It is not unrealistic to think that future legislators may enact unjust laws. (Indeed, even at present, it is quite likely that we have certain criminal laws which society will, in the future, recognize as unjust. Throughout history, there have always been certain norms in whose soundness people once had great confidence, but which have been overturned as society progressed.) Legislators need not even be thoroughly corrupt for such laws to be created – they may be well intentioned but misguided. If society were already well accustomed to the authorities' use of mandatory neurointerventions to rehabilitate justly convicted offenders, it would be that much easier for the authorities forcibly to impose such interventions on those convicted under unjust laws. We need to set limits on the state's power by ensuring that there is some sphere of individual liberty into which the state may never intrude. Whatever other liberties one may potentially "forfeit," an individual should, at least, be able to veto any biomedical interference with her own brain.

A Retributive Objection

The fact that state can force offenders to undergo traditional punishments does not imply that it has the right to force offenders to undergo neurointerventions (Cf. Ryberg and Petersen 2013a). On a retributive view, the state's right to sentence offenders to prison, community service, etc. stems from its right to punish. On Michael Moore's (1997) influential retributive theory, punishment is the intentional infliction of deserved suffering on moral wrongdoers, on the basis that such suffering is intrinsically good. However, most theorists agree that neurointerventions should not be viewed as punishment in the retributive sense. Neurointerventions are methods of rehabilitation, intended to protect society and/or to help offenders themselves to lead better lives. As Bomann-Larsen (2013) points out, to view mandatory neurointerventions as a form of retributive punishment would involve abandoning fundamental principles of justice such as the principle that punishment must not be "cruel or unusual." It would seem the height of cruelty to impose neurointerventions because of the suffering they could produce (e.g., the distress of having one's mind manipulated against one's will and any side effects these interventions might have).

As noted above, the most plausible reasons for employing neurointerventions are for the protection of others and/or to benefit the offender himself. There is a strong argument that the state's power to intervene for such reasons must be constrained by the need to respect the individual's status as an autonomous agent. If there were no such constraint, then the state could sacrifice individuals whenever this promoted the general welfare, or whenever *the state* judged this to be in the individual's best interests. Forcing an offender to undergo neurointerventions, without his

consent, would disregard his rational capacities and would arguably involve treating him as an object, as a being without human dignity (See section "Mandatory Neurointerventions Objectify Offenders", below). In contrast, a retributivist might argue that the state's right to force offenders to undergo traditional punishment already respects their autonomy, since it is intended as the deserved penalty for an autonomous action (not as a means of pre-empting future autonomous decisions to reoffend) and thus does not require explicit consent.

A Non-Retributive Objection

This section will now sketch an objection to mandatory neurointerventions that is very different from the retributive objection discussed earlier. The alternative line of argument that will now be developed should appeal to some of those who reject the idea that the suffering of the guilty is intrinsically good. Derk Pereboom (2001), for instance, believes that offenders' interests and preferences are of equal moral weight to anyone else's. Yet, he obviously acknowledges that serial rapists and murderers cannot simply be allowed to roam free. He does not claim that society has a general, unlimited right to sacrifice individuals for the overall welfare. Rather, he argues that society does have a right (similar to the right of self-defense) to protect itself from individuals who pose a serious threat to others. This right is subject to various constraints including a necessity constraint – the state should set back dangerous individuals' interests no more than is necessary to prevent them from harming others. He draws an analogy with quarantine. If society has the right to protect itself against carriers of dangerous diseases by subjecting them to quarantine, then it has the right to protect itself from dangerous offenders by incapacitating them. He argues that (like quarantine) the conditions of an offender's incapacitation should be humane and should still give weight to the offender's preferences. Without effective neurointerventions, constraining offenders' movements would often be the only humane and effective way of protecting society. In such circumstances, given society's right to protect itself, mandatory detention would be legitimate. However, Pereboom's view seems to imply that if safe and effective neurointerventions were developed, then society would lose the right simply to impose mandatory detention. Nor would it gain the right to impose mandatory neuroenhancement. Rather, it would show most respect for offenders' equal moral status to give them the choice between detention until they become non-dangerous, and, alternatively, neuroenhancement.

Thus, whether one takes a retributive position (like Moore's) or a radically non-retributive position (like Pereboom's), there are strong reasons, from within these perspectives, to oppose mandatory neuroenhancement of offenders.

Mandatory Neurointerventions Objectify Offenders

There are good reasons to resist any tendency to objectify offenders (reasons that are available to theorists regardless of whether they view offenders' moral status

from a retributive or a non-retributive standpoint). Offenders are already among the most despised members of society. People's desire to see offenders suffer often exceeds anything that could be morally justified, even on the most severe retributive view. There is a danger that feelings of disgust for criminals could lead society to lose sight of their humanity. It is therefore particularly important to have clear restrictions on the ways in which the state can treat them. The term "objectification" can be used in different ways. The conception of objectification that this chapter adopts is influenced by sociological discussions of the ways in which disfavored groups within society have historically been objectified or treated as "subhuman" (see, e.g., Reicher 2006). This kind of objectification typically involves creating a division between "them" and "us" that excludes the objectified group (cf. Shaw 2012). If offenders, unlike the rest of society, are subject to mandatory neurointerventions, then they are forced to relinquish something that many people regard as fundamental to their own status as persons – their control over their mental and bodily integrity. This would radically set criminals apart from the rest of society – leading us to categorize them as "the other" – as beings who lack basic rights and are largely outside the sphere of our moral concern. True, some of our basic moral rights, e.g., freedom of movement, are qualified. The right to freedom of movement does not provide immunity from state interference after one has committed a serious crime. However, there is a categorical difference between interfering with an offender's freedom of movement and forcibly interfering with the offender's mental and bodily integrity. Our right to mental and bodily integrity is more fundamental (and essential to human dignity) and merits even greater protection than our right to freedom of movement. Offenders cannot simply be said to "forfeit" their right to mental and bodily integrity without giving up something that is central to their status as persons.

The Voluntary Use of Neuroenhancement in Rehabilitation Programs

Informed Consent

It is important that offenders are given full information about the potential side effects and efficacy of neurointerventions. However, as Henry Greely (2008, 2012) points out, gathering such information may be far from straightforward. It is difficult to draw conclusions from animal models about the effects of brain interventions, since the human brain is very different even from the brains of closely related species, e.g., chimpanzees, let alone the brains of lab rats. Experiments on primates are also extremely ethically controversial. Usually, when a medical intervention is developed in order to treat a disease, information about its effects can be gathered from trials on volunteers who participate in the hope of benefitting from the treatment. However, since neuroenhancements are given to address antisocial behavior (i.e., largely for the benefit of people other than the recipient), volunteers may be much harder to find.

However, if an intervention for controlling offending behavior also has another clinical use, then a certain amount of data on side effects may already exist. For instance, testicular pulpectomy has been employed in certain countries (e.g., Germany) to help reduce sexual reoffending (see McMillan 2013). Some information about the side effects of this procedure is available since it is also an established treatment for prostate cancer. As noted above, SSRIs and propranolol may reduce violent and racist tendencies respectively. SSRIs are also used as antidepressants and propranolol is a treatment for hypertension and their use in this context has generated information about their side effects.

It might be thought that, in order to make a truly informed choice, the offender should also be provided with information about the risks of refusing treatment (e.g., statistical information about the mental health risks of remaining for a potentially longer period in prison). However, some might worry that if officials emphasized the risks of prison when offering neurointerventions, the state might be perceived as exploiting the fear of prison to "coerce" offenders into accepting neurointerventions (cf. McMillan 2013).

Voluntary Consent

If an offender is offered neuroenhancement as a condition of early release from prison, then he is faced with a choice between unappealing alternatives. However, this does not seem to preclude the possibility of giving voluntary consent to neurointerventions, provided that the offender is not dominated by a literally irresistible fear of spending a longer time in prison (Bomann-Larsen 2013). After all, patients are often considered capable of freely consenting, despite being faced with very hard choices, e.g., a choice between undergoing risky and disfiguring surgery to remove diseased tissue and, alternatively, refusing treatment and dying from their disease.

Threats and Inappropriate Offers

In order for consent to be normatively valid, more is required than minimal voluntariness (i.e., the absence of irresistible impulses) and adequate information. A valid consent to an offer of a biomedical intervention might plausibly be thought to have the following moral and legal effects: if such consent is given, then the person making the offer ("the offerer") has not wronged the person receiving the intervention ("the recipient") by providing that intervention. If the offerer were accused of having wronged the recipient, then the former could rely on the latter's consent as a defense. When thinking about whether valid consent exists, it is common to focus on the freedom of the person giving consent, i.e., the subjective sense of pressure she is experiencing and how restricted her alternatives are. However, if consent is to have the moral and legal effects just described, then it also makes sense to look at the conduct of the person seeking consent. It is well

accepted that a person's "consent" can be invalid if it were obtained by (nontrivial) threats to violate that person's rights. This is not because a person who is subject to such threats necessarily faces a harder choice or feels more pressure than the dying patient in the example given above. The victim's consent is invalidated because of the wrongful conduct of the party threatening to violate the victim's rights. The issuing of a threat of this kind seriously wrongs the victim and therefore, the activity that is "consented" to, as a result of that threat, also wrongs the victim. Bublitz and Merkel make a similar argument, rightly emphasising that a subjective, psychological sense of pressure will not automatically invalidate consent (Bublitz and Merkel 2009, 2013). However, they seem to endorse Feinberg's claim that wrongful threats invalidate consent because threats render consent involuntary (Bublitz and Merkel 2009, pp. 372–373; Feinberg 1987, p. 196). In contrast, I favor the view that wrongful threats/offers can invalidate a person's consent, even if the consent is voluntary. Voluntariness is one necessary condition for the validity of consent to a neurointervention, but the moral permissibility of the proposal to intervene is plausibly another, separate, requirement for the validity of consent (see Bomann-Larsen 2013). It is the latter requirement that is breached by wrongful threats.

If an offender were offered a neurointervention as a condition of early release from prison, it would be a mistake to construe his original prison sentence as a "threat" to violate the offender's rights. If this original sentence is a just response to his crime, then by definition, it does not violate his rights (Bomann-Larsen 2013). The offer of a neurointervention is a genuine offer. However, Bomann-Larsen argues that certain offers of medical interventions can also seriously wrong the other party and thus invalidate consent to such interventions. She cites Feinberg's (1986) example of a prison governor who offers to commute the sentence of a prisoner on death row if the prisoner participates in a medical experiment. The governor is abusing his position – he has no right to propose that the prisoner participate in such an experiment. He is taking unfair advantage of the prisoner's predicament. His offer wrongs the prisoner and arguably invalidates the prisoner's consent to the experiment. Note again that the key point is not the hard choice that the offender faces. In the example just given, the offender's alternatives are both very unappealing – death on the one hand and a potentially risky experiment on the other. However, as Bomann-Larsen points out, the governor's offer would still be wrongful if he presented the prisoner with an easier choice, e.g., "I will spare your life if you bring me tea every morning for 10 years." This offer is still wrongful, since the governor, qua governor, is not entitled to propose that the prisoner becomes his servant.

Similarly, Bomann-Larsen persuasively argues that the state can wrong an offender by offering him neurointerventions that are aimed at altering personality traits or private behavior that the state has no right to interfere with. The offender may be an untrustworthy friend, an unfaithful spouse, and an uncooperative colleague. But such traits or behavior are none of the state's business. In contrast, the state is entitled to seek to fix the problematic behavior for which the offender was convicted. Bomann-Larsen argues that, ideally, any intervention offered to offenders should be strictly necessary to address that behavior.

Ryberg and Petersen (2013a) raise the following objection to this necessity constraint. They point out that the state has a general duty to protect the public from all crimes, not just the one for which the offender was convicted. If, after the offender is convicted of one crime, a psychological evaluation reveals that he has a disposition to commit another type of crime, then they suggest that the state may have the right to offer him a neurointervention to address this latter disposition. In response, it is submitted that the state's duty to tackle criminal behavior must be subject to various procedural constraints. For instance, the state cannot subject an individual to punishment for a crime unless it is proved beyond reasonable doubt that he committed it. One justification for the beyond reasonable doubt standard is based on the kind of trust and respect the state should show to citizens as members of the moral community (Duff 2001). Rather than the default position being an attitude of suspicion, the state's relationship with citizens should begin with a strong presumption that citizens are morally well motivated (or at least that they do not violate the central moral norms underlying the criminal law). It should take strong evidence to rebut this presumption. Establishing beyond reasonable doubt that the offender has already committed a criminal act provides fairly strong evidence that the offender has a disposition to commit acts of this kind. Thus, it does not show disrespect, or distrust, to offer him interventions to address this disposition. However, it would be inappropriate to assume that the offender were liable to commit a range of other, unrelated crimes on the basis a psychological evaluation, rather than past behavior. The offender's conviction for a particular crime is not meant to be the starting point for a general investigation into his character and dispositions to commit unrelated crimes. That would also present too great an intrusion into individual liberty.

Conclusion

This chapter has critically discussed some of the main arguments for prohibiting the use of neurointerventions in offender rehabilitation programs and for the opposite position of making such interventions mandatory. Both of these extreme positions are ultimately unconvincing. If sufficiently safe and effective neurological techniques for addressing criminal behavior were developed, the best policy would be to offer them on a voluntary basis. In order for consent to such interventions to be genuine and valid, certain conditions must be met. Consent must, of course, be informed and voluntary. In addition, it is important to remember that the concept of "valid consent" (in a robust, morally significant sense) is closely bound up with the idea of respect. If the state is to cite the offender's consent as part of the moral justification for intervening in his brain, then it must ensure that when obtaining that consent, it did not act in an exploitative manner or overstep its authority, but instead that it respected the offender's status as a moral agent.

Future Directions

The development of neurointerventions with the potential to address offending behaviour is still in its early stages. More research needs to be done on the side effects and effectiveness of such interventions, including possible long-term side-effects. Ethicists will be a better position to deliberate about the acceptability of offering neurointerventions when the risks attached to such interventions are properly understood.

Cross-References

▶ Real-Time Functional Magnetic Resonance Imaging–Brain-Computer Interfacing in the Assessment and Treatment of Psychopathy: Potential and Challenges
▶ The Morality of Moral Neuroenhancement

References

Adebowale, V. (2010). Diversion not detention. *Public Policy Research, 17*(2), 71–74.

Blair, J., Mitchell, D., & Blair, K. (2005). *The psychopath: Emotion and the brain*. Oxford: Blackwell.

Bomann-Larsen, L. (2013). Voluntary rehabilitation? On neurotechnological behavioural treatment, valid consent and (in)appropriate offers. *Neuroethics*, *6*, 65–77.

Bublitz, J., & Merkel, R. (2013). Guilty minds in washed brains? Manipulation cases, excuses and the normative prerequisites of liberal legal order. In N. Vincent (Ed.), *Neuroscience and legal responsibility* (pp. 335–374). Oxford: OUP.

Crockett, M., Clark, L., Hauser, M., & Robins, T. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Psychological and Cognitive Science, 107*, 17433–17438.

Cunningham, W., Johnson, M., Raye, C., Gatenby, J., Gore, J., & Banaji, M. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15*, 806–813.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*(3), 228–245.

Douglas, T. (2013a). Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics, 27*(3), 160–168.

Douglas, T. (2013b). The relationship between effort and moral worth: Three amendments to Sorensen's model. *Ethical Theory and Moral Practice*. Online First. doi:10.1007/s10677-013-9441-4.

Duff, R. (2001). *Punishment, communication and community*. Oxford: OUP.

Farah, M. (2004). Emerging ethical issues in neuroscience. *Nature Neuroscience, 5*(11), 1123–1130.

Feinberg, J. (1986). *Harm to self* (The moral limits of the criminal law). New York: Oxford University Press.

Ferari, P., et al. (2005). Escalated aggressive behavior: Dopamine, serotonin and GABA. *European Journal of Pharmacology, 526*, 51–64.

Freedman, C. (2000). Aspirin for the mind? Some ethical worries about psycho-pharmacology. In E. Parens (Ed.), *Enhancing human traits: Ethical and social implications* (pp. 135–150). Washington, DC: Georgetown University Press.

Greely, H. (2008). Neuroscience and criminal justice: Not responsibility but treatment. *Kansas Law Review, 56*, 1103.

Greely, H. (2012). Direct brain interventions to 'treat' disfavored human behaviors: Ethical and social issues. *Clinical Pharmacology & Therapeutics, 91*(2), 163–165.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*(2), 102–111.

Harrison, K., & Rainey, B. (2013). *The Wiley-Blackwell handbook of legal and ethical aspects of sex offender treatment and management*. Oxford: Wiley-Blackwell.

Hart, A., et al. (2000). Differential response in the human amygdala to racial outgroup vs. ingroup face stimuli. *Neuroreport, 11*, 2351–2355.

Levy, N. (2007). Neuroethics: Challenges for the 21st century. Cambridge: Cambridge University Press.

Matthews, E. (2007). *Body-subjects and disordered minds. Treating the whole person in psychiatry*. Oxford: OUP.

McMillan, J. (2013). The kindest cut? Surgical castration, sex offenders and coercive offers. *Journal of Medical Ethics*. Published [199] Online First: 11 May 2013. doi:10.1136/200 medethics-2012-101030.

Mill, J. (1909). *On liberty*. New York: P.F. Collier and Son, chapter 5.

Ministry of Justice. (2013). *Proven re-offending statistics, quarterly bulletin*. October 2010–September 2011. England and Wales. https://www.gov.uk/government/publications/proven-re-offending–2

Moore, M. (1997). *Placing blame*. Oxford: Oxford University Press.

Newman, J., et al. (2010). Attention moderates the fearlessness of psychopathic of-fenders. *Biological Psychiatry, 67*, 66–70.

Olsen, J. (2006). Depression, SSRIs, and the supposed obligation to suffer mental-ly. *Kennedy Institute of Ethics Journal, 16*(3), 283–303.

Penney, S. (2012). Impulse control and criminal responsibility: Lessons from neu-roscience. *International Journal of Law and Psychiatry, 35*, 99–103.

Pereboom, D. (2001). *Living without free will*. Cambridge: CUP.

Phelps, E., et al. (2000). Performance on indirect measures of race evaluation pre-dicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729–738.

Rakic, V. (2013). Voluntary moral enhancement and the survival-at-any-cost bias. *Journal of Medical Ethics*. Online First. doi:10.1136/medethics-2012-100700.

Rawls, J. (2005). *A theory of justice*. Cambridge, MA: Harvard University Press.

Reicher, S. (2006). Saving Bulgaria's Jews: An analysis of social identity and the mobilisation of social solidarity. *European Journal of Social Psychology*, *36*, 49–72.

Ryberg, J., & Petersen, T. (2013a). Neurotechnological behavioural treatment of criminal offenders – A comment on Bomann-Larsen. *Neuroethics, 6*, 79–83.

Ryberg, J., & Petersen, T. (2013b). Surgical castration, coercion and ethics. *Journal of Medical Ethics*. Published Online First: 19 June 2013b. doi:10.1136/medethics-2013-101508.

Savulescu, J., & Persson, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy, 25*, 162–177.

Shaw, E. (2011). Free will, punishment and neurotechnologies. In B. Van den Berg & L. Klaming (Eds.), *Technologies on the stand. Legal and ethical ques-tions in neuroscience and robotics* (pp. 177–194). Nijmegen: Wolf Legal Publishers.

Shaw, E. (2012). Direct brain interventions and responsibility enhancement. *Criminal Law and Philosophy*. Online First. doi:10.1007/s11572-012-9152-2.

Sorensen, K. (2010). Effort and moral worth. *Ethical Theory and Moral Practice, 13*(1), 89–109.

Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Cowen, P., & Hewstone, M. (2012). Propranolol reduces implicit negative racial bias. *Psychopharmacology, 222*, 419–424.

Timpe, K. (2011). Tracing and the epistemic condition on moral responsibility. *The Modern Schoolman, 88*, 5–28.
Vilhauer, B. (2009). Free will skepticism and personhood as a desert base. *Canadian Journal of Philosophy, 39*(3), 489–511.
Wellman, C. (2012). The rights forfeiture theory of punishment. *Ethics, 122*, 371–393.

Cases

Loving v Virginia 388 US 1 (1967)

Section XVIII

Feminist Neuroethics

Feminist Neuroethics: Introduction

87

Peggy DesAutels

Contents

Abstract

Feminist neuroethics uses gender as a lens to approach neuroethics, drawing on an ever-growing body of feminist theory and scholarship. This feminist section of the *Handbook of Neuroethics* contains four articles. The first is by Robyn Bluhm on ▶ "Feminist Philosophy of Science and Neuroethics" (Chap. 88). Bluhm maintains that feminist philosophy of science can provide theoretical and methodological resources to address questions posed within neuroethics. The second article in this section is by Peggy DesAutels on ▶ "Feminist Ethics and Neuroethics" (Chap. 89). She examines if and to what degree there are sex/gender differences in moral judgments, behaviors, and traits and concludes that most differences are attributable not to sex-based brain differences between men and women, but to distinct and persistent types of psychologies found in members of oppressing and oppressed groups. The third article is by Anne Jaap Jacobson entitled ▶ "A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience" (Chap. 90). Jacobson examines implicit bias and the degree to which it can be changed. Her focus is on bigoted racial bias, but she is motivated by concerns for gender bias as well. The fourth and final article in this

P. DesAutels
Department of Philosophy, University of Dayton, Dayton, OH, USA
e-mail: peggy.desautels@gmail.com; Peggy.DesAutels@notes.udayton.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_80,

section is by Cordelia Fine and Fiona Fidler on ▶ "Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform" (Chap. 91). They describe how the current statistical method of null hypothesis significance testing specifically contributes to scientific error in sex/gender neuroscience. They suggest that sex/gender neuroscience may therefore provide a valuable model to motivate, on ethical grounds, statistical reform within the psychological sciences.

Characteristics

Feminist neuroethics uses gender as a lens to approach neuroethics, drawing on an ever-growing body of feminist theory and scholarship. Although there are many types of feminists and feminist theories, all consider gender to be an important analytic lens. All promote political, economic, and social equality of the sexes. And all agree that human beings live gendered lives in a social world that contains both explicit and implicit misogyny, sexism, and androcentrism. A feminist expectation is that addressing gender-based injustices will result in a better world for both women and men. Feminist theory often draws on and overlaps with the critical theories of other oppressed groups marked by, e.g., race, class, disability, or sexual orientation. The shared goal of such critical theories is a society free of oppression, bigotry, and group-based harms of all sorts.

There are several feminist concerns related to neuroethics. One is the degree to which neuroscientists treat the male brain as the "standard" brain and the female brain as the "other." Another is the potential for sex/gender bias, if not all-out sexism, in sex-difference research and how best to promote research that is less biased. Some feminists have specific concerns over some neuroscientists' claims that there are innate sex-based differences in our psychologies. And a broader set of questions of possible concern to feminist scholars includes the following: To what degree can or should scientific research including neuroscientific research be "value-neutral" and "objective"? What research questions are and are not posed, and why? What hypotheses are and are not tested? Whose interests are and are not promoted by various neuroscientific research programs?

Outstanding books that take a feminist approach to neuroethics include *Sexing the Body*: *Gender Politics and the Construction of Sexuality* by Anne Fausto-Sterling (2000), *Delusions of Gender*: *How Our Minds*, *Society*, *and Neurosexism Create Difference* by Cordelia Fine (2010), *Brain Storm*: *The Flaws in the Science of Sex Differences* by Rebecca M. Jordan-Young (2010), and *Neurofeminism*: *Issues at the Intersection of Feminist Theory and Cognitive Science* edited by Bluhm et al. (2012).

This feminist section of the *Handbook of Neuroethics* contains four articles. The first is by Robyn Bluhm on ▶ "Feminist Philosophy of Science and Neuroethics" (Chap. 88). Bluhm maintains that feminist philosophy of science can provide theoretical and methodological resources to address questions posed within neuroethics. These resources can be applied to contemporary research on sex/gender differences as well as to many other areas of neuroscience under ethical

scrutiny. After briefly surveying recent feminist critiques of neuroscience research on sex/gender differences, she demonstrates how two main themes in feminist philosophy of science – feminist standpoint theory and feminist empiricism – drew on this research to develop new theoretical approaches to philosophy of science. Finally, she examines how some of the key features of feminist philosophy of science might apply to research in neuroscience.

The second article in this section is by Peggy DesAutels on ► "Feminist Ethics and Neuroethics" (Chap. 89). DesAutels focuses on the branch of neuroethics devoted to the neuroscience of moral judgments and behaviors. She takes a feminist approach to this particular branch of neuroethics. She examines if and to what degree there are sex/gender differences in moral judgments, behaviors, and traits and asks: If there are apparent differences, what are the most likely causes? She argues that most differences are attributable not to sex-based brain differences between men and women, but to distinct and persistent types of psychologies found in members of oppressing and oppressed groups.

The third article is by Anne Jaap Jacobson entitled ► "A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience" (Chap. 90). Jacobson examines implicit bias and the degree to which it can be changed. Her focus is on bigoted racial bias, but she is motivated by concerns for gender bias as well. She draws on critical race theory and neuroscientific investigations of biased judgments to address the questions: (1) To what degree are those who hold harmful implicit biases morally responsible for their biased actions and attitudes? (2) To what degree and using what strategies can harmful implicit biases be mitigated?

The fourth and final article in this section is by Cordelia Fine and Fiona Fidler on ► "Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform" (Chap. 91). In this article, they briefly review the benefits of the estimation statistical approach as a means to producing reliable information about nature. They then describe how the current statistical method of null hypothesis significance testing specifically contributes to scientific error in sex/gender neuroscience. They highlight the potential social harm that can arise from such errors in this area of research and suggest that sex/gender neuroscience may therefore provide a valuable model to motivate, on ethical grounds, statistical reform within the psychological sciences.

Cross-References

- ► A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience
- ► Brain Research on Morality and Cognition
- ► Developmental Neuroethics
- ► Feminist Ethics and Neuroethics
- ► Feminist Philosophy of Science and Neuroethics
- ► Moral Cognition: Introduction
- ► Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences
- ► No Excuses: Performance Mistakes in Morality

► Psychology and the Aims of Normative Ethics
► Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform

References

Bluhm, R., Jacobson, A. J., & LeneMaibom, H. (Eds.). (2012). *Neurofeminism: Issues at the intersection of feminist theory and cognitive science*. New York: Palgrave Macmillan.

Fausto-Sterling, A. (2000). *Sexing the body: Gender politics and the construction of sexuality*. New York: Basic Books.

Fine, C. (2010). *Delusions of gender: How Our minds, society, and neurosexism create difference*. New York: W.W. Norton & Company.

Jordan-Young, R. M. (2010). *Brainstorm: The flaws in the science of sex differences*. Boston, MA: Harvard University Press.

88 Feminist Philosophy of Science and Neuroethics

Robyn Bluhm

Contents

Abstract

As a recognized subdiscipline of applied ethics, neuroethics is quite new; however, ethical analyses and critiques of neuroscience have a long history. A significant proportion of these critiques have come from feminists, in response to research that aimed to uncover differences in the brains of women and men and often to use these differences to support political claims about women's place in society. This entry surveys these feminist critiques and looks at the potential for recent work in feminist philosophy of science to contribute to neuroethics. Because it addresses the appropriate role for values in scientific research, feminist philosophy of science can provide useful tools for the analysis of sex/gender difference research and also for a variety of other topics of interest to neuroethicists.

R. Bluhm
Department of Philosophy and Religious Studies, Old Dominion University, Norfolk, VA, USA
e-mail: rbluhm@odu.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_158,

Introduction

As a recognized subdiscipline of applied ethics, neuroethics is quite new; however, ethical analyses and critiques of neuroscience have a long history. A significant proportion of these critiques have come from feminist scholars, in part because of (and focusing on) neuroscience research on sex/gender differences. Yet, to date, there have been relatively few attempts to develop an explicitly feminist approach to neuroethics. Peggy DesAutels (2010) has called for neuroethics to pay more attention to feminist work on sex difference research. Both Deboleena Roy (2012) and Katrin Nikoleyczik (2012) have shown that work by feminist science studies scholars is very relevant to neuroethics. This paper does the same for feminist philosophy of science in the analytic tradition. Feminist philosophy has been an active area of research for nearly three decades, and it was inspired, in part, by feminist work that criticized sex/gender difference research in neuroscience and psychology. The field has now developed to the point where it can provide theoretical and methodological resources useful for analyzing questions addressed by neuroethics. These resources can be applied to contemporary research on sex/gender differences but also apply equally well to many other areas of neuroscience that have prompted ethical scrutiny.

This paper will begin with a brief survey of some feminist research that critically assesses neuroscience research on sex/gender differences. (As will be discussed later in the paper, there is debate over the extent to which differences between women and men should be attributed to sex or to gender. The compound term indicates that there is no way to settle this debate (see also Kaiser (2012)) It then shows how two main themes in feminist philosophy of science – feminist standpoint theory and feminist empiricism – drew on this research but went well beyond these critiques to develop new theoretical approaches to philosophy of science. Although these two forms of feminist philosophy were originally in tension with each other, they have always shared much common ground and have converged over the years. The final section of the paper highlights some of the key features of feminist philosophy of science and show how they might apply to research in neuroscience.

Feminist Critiques of Neuroscience

Feminist scientists and science studies scholars have documented numerous cases in which prevailing concepts of the nature of women and of men have shaped scientific research, particularly in biology and the social sciences. Because neuroscience promises to explain so much about our characters and behaviors, it has been a key area of research on sex/gender differences, attractive to those who sought to explain – and sometimes to justify – the different places in society occupied by women and by men. This section briefly describes some of this research and shows that it raises important questions for philosophy of science.

While biological explanations of women's inferiority to men date back at least to Aristotle's work (see Tuana (1993) for a discussion of this history), the second half of the nineteenth century saw a large increase in research on sex differences. It also

marked the "first wave" of feminism in the United Kingdom and the United States. Women were demanding the right to vote and, more generally, to play a greater role in public life. Their political activity threatened the existing social order, which was based on a sharp distinction between the roles of women and of men; while men were appropriately concerned with politics, commerce, and science, women's place was in the home.

The different roles of women and men had largely been taken for granted as the being result of natural differences in the characteristics and abilities of each sex. Now that the roles were being challenged, however, it became important to demonstrate that they really were the result of such natural differences. Scientists began to investigate the differences between women and men. Janet Sayers surveys research in the late nineteenth century that was undertaken specifically in order to demonstrate that, owing to women's frailty, higher education would damage their health and fertility, thus also threatening the future of the human species (Sayers 1982, Chap. 2).

Elizabeth Fee has traced the attempts of craniologists to show that women's intellectual inferiority was the result of differences between their brains and men's. Craniologists, as their name suggests, actually measured the skull, rather than the brain, but contemporary theory said that because the skull "formed a faithful cast of the underlying brain. . .measurement of crania could therefore be substituted for direct measurement of the brain" (Fee 1979, p. 420). Demonstrating that women's skulls were smaller would therefore provide a biological explanation for their inferior intelligence and thus both explain and justify their exclusion from public life.

This project, however, proved to be more difficult than anticipated. Fee documents the many measurement techniques that craniologists used to try to demonstrate the inferiority of female skulls and the equally many failures to find a clear sex difference. Scientists looked for absolute size differences, size differences relative to overall body size, and differences in the relative size and shape of various parts of the skull. All of these approaches failed to show the difference that scientists sought. Yet scientists never questioned the hypothesis that they were attempting to prove that women's skulls would provide evidence of their intellectual inferiority. Nor did they question the more general hypothesis that women were less intelligent than men. Fee notes that "the whole enterprise thrived despite annoying setbacks; as one [measurement] index failed to measure up to expectations, another appeared, claiming to hold the solution all earlier problems" (p. 427). Eventually, the entire science of craniology "collapsed under its own weight" (p. 432) and was abandoned.

Fee's research on this era was conducted in the 1970s, a time at which second-wave feminism was leading to women's increased participation in the workplace and fewer people would overtly espouse the belief that women are intellectually inferior to men. Yet beliefs in the different strengths and abilities of women and men were still common, and still used to justify women's exclusion from certain fields, and once again scientists were busy trying to find a biological basis for those differences. Just as Fee's historical research uncovered sexism in nineteenth-century scientists, work by her contemporaries in the biological and social scientists showed that

similar assumptions about differences between women and men informed scientific work on sex differences in the 1970s and 1980s.

In fact, this period was an especially busy one for neuroscientists interested in sex/gender differences. In her critique of biological theories of sex differences, Ruth Bleier links this phenomenon directly to social changes, noting that "[a]s social movements threaten the social order, it is a recurrent phenomenon that corresponding scientific theories emerge that implicitly defend the *status quo*" (1984, p. vii). Like the nineteenth-century craniologists, Bleier's contemporaries were informed by the social background within which their research was conducted.

In some ways, the neuroscience research conducted in the middle of the twentieth century is shockingly similar to that of the previous century. During the 1980s, there was a spate of research examining the possibility that there were sex differences in the corpus callosum, a bundle of nerve fibers connecting the two hemispheres of the brain. These differences were hypothesized on the basis of a then-current theory that suggested that men's brains were more "lateralized" than those of women so that the left and right hemispheres operated relatively independently. Since women's brains were more closely connected, the hemispheres operated in tandem. Differences in lateralization were supposed to explain the (supposedly) different strengths of women and men, such as women's greater verbal ability and men's superior visuo-spatial skills. Anne Fausto-Sterling surveys this research on the corpus callosum and finds that it is in many ways similar to the craniology research of the previous century. Because this brain structure has a complex shape, there are many possible ways to measure it and scientists appear to have tried them all, yet "no matter how they carve up the shape, only a few researchers find absolute sex differences in CC area. A small number report that males and females have differently shaped corpus callosums. . ., even though the shape does not translate into a size (area or volume) difference" (2000, pp. 130–1). Yet scientists continued to debate whether there was a difference. Fausto-Sterling concludes that the longevity of the debate "speaks to how entrenched their expectations about biological differences remain" (p. 145).

Despite the similar methods and tactics used in both of the eras of research I have described, there are also important differences between them. One such difference is that the more recent period included an explicit debate about the "naturalness" of any differences found between the sexes. While in the nineteenth century, this was taken for granted, it is now hotly contested. Bleier ties this debate to political changes: "In the face of the contemporary vigorous resurgence of the women's movement and feminist scholarship, biological determinist theories have become a conspicuous part of our scientific and popular cultures" (1984, p. vii). Similarly, Sayers (1982) and Birke (1986) provide examples of the use of scientific research to defend conservative political positions. In general terms, biological determinism says that what we are is an inevitable result of our biology. When it comes to sex/gender differences, the implication of biological determinism (which, as noted above, is sometimes made explicit) is that the different social positions occupied by men and women are a direct and inevitable result of biological differences.

One key aspect of feminist thought during this period was the idea that the biological and psychological differences between women and men were

not inseparable. The former were assigned to the category of "sex" and were acknowledged to be natural, biological differences, but the latter – reflecting the different traits and abilities associated with each sex – were the result of the social environment. In other words, physical differences between women and men were thought to be irrelevant to differences in psychological and behavioral characteristics. Rather, these are the result of being raised in a particular social environment that included expectations about appropriate behavior for each sex. These latter characteristics were subsumed under the label of "gender" as opposed to sex. Distinguishing between biology and society, between sex and gender, was intended to undercut biological determinism. As Anne Fausto-Sterling explains, feminists distinguishing between sex and gender believed that "it was social institutions, themselves designed to perpetuate gender inequality, that produced most of the differences between men and women" (2000, p. 638).

Yet almost as soon as it was established, the distinction between sex and gender came under attack, not just by biological determinists but also by feminists who questioned the "naturalness" of physical, sex differences, arguing that they, too, were at least in part a product of differences in the way that girls and boys were raised and women and men lived. For example, differences in muscle mass and development were only partly the result of different hormone levels. They also occurred because boys were encouraged to be active and play sports, while girls were taught to prefer quieter, more sedentary games.

The distinction between innate influences on biology and the effects of experience was recognized as being particularly unclear when it came to the brain. Anatomical differences, for example, those in the corpus callosum, could be the result of a hardwired developmental program, as the biological determinists would have it. But it was also possible that the environment played an important role in shaping the structure of the brain. Developmental neurobiologists were learning how much of brain development occurred after birth and just how responsive the developing brain was to experience. Since boys and girls begin to have different kinds of experience from birth, it is not surprising that their brains develop differently. Nor were these arguments dependent merely on a vague appeal to cultural factors; accounts of environmentally influenced brain development drew heavily on neuroscience research. Anne Fausto-Sterling draws on developmental systems theory to outline the ways in which biology and the social environment might interact. She notes that merely *finding* biological differences in the brains of women and men does not show that those differences are an inevitable result of different (innate) developmental trajectories. They also are likely to reflect the different social environments in which girls and boys are raised. We know that there is a great deal of postnatal development of the brain and that this development is sensitive to the environment, so it is not unlikely that differences in the social environment will affect the brain. Fausto-Sterling draws the conclusion that "if differential social experience produces differences in brain anatomy and thus in brain function, later experiences would then be interpreted and integrated by a differently functioning brain" (p. 125).

One possible response to disagreements about the relative contributions of "nature" and "nurture" to observed differences in the brains of women and of men might be that we need to collect more data, that eventually scientists can amass evidence that conclusively supports one side or the other of the debate. This turns out not to be the case; one of the key insights of philosophy of science has been that any amount of data is always consistent with more than one theoretical explanation. The problem of underdetermination, as this point is called, means that scientists need to go beyond the data in choosing one theory over another. This does not mean that "anything goes" or that it is never possible to justify the choice of one theory over another. It simply means that scientists always draw on other considerations when interpreting data or developing theoretical explanations of phenomena. One type of consideration that informs their choices is epistemic values such as simplicity, explanatory breadth, and consistency with other theories in related domains. The second type of consideration, which is often only an implicit influence on scientists' choices, is background assumptions about the nature of the phenomenon being studied. In the case of research on sex/gender differences, background assumptions about the extent to which brain development is responsive to the environment influence the interpretation of experimental data. For example, recall the research discussed above on sex differences in the corpus callosum. Bell and Variend (1985) conducted a study in which they examined this structure in young children and found no sex/gender differences, despite the fact that some researchers had found such differences in adult brains. Those who are inclined to believe that sex/gender differences in the brain are largely a result of environmental differences would point to this study as evidence for their claims. But those who think that sex/gender differences are biologically determined would say that the differences that eventually emerge are the result of distinct (innate) developmental programs for each sex, something like the changes that emerge at puberty. Background assumptions about the source of sex/gender differences influence the way that people interpret the available evidence, biasing them to one interpretation or the other.

But science is not supposed to be biased; it has traditionally been assumed. Good science is thought to be objective and impartial. After all, part of the criticism leveled against neuroscientists by the feminist scholars discussed above was that their sexist beliefs, or at least their commitment to the existence of sex differences (whatever their origin), had a negative influence on their science. And part of the criticism raised against feminist science and feminist science studies is that "wishful thinking" about equality or background commitments to gender equality would lead to feminists' ignoring evidence that did not support their views.

It is here that feminist philosophy of science can make an important contribution to neuroethics; one of the main concerns of this area of research has been to clarify the ways in which science is – and should be – influenced by values. The next section provides some background about the origins and development of feminist philosophy of science and then turns to some of its key insights.

The Development of Contemporary Feminist Philosophy of Science

Much of the feminist criticism of neuroscience described above was conducted by feminist scientists who were themselves active researchers in the fields they critiqued. They therefore had a detailed knowledge of the methods and theories they discussed. In her discussion of emerging work in feminist philosophy of science, Sandra Harding calls their work "spontaneous feminist empiricism," which "arose as the 'spontaneous consciousness' of feminist researchers in biology and the social scientists who were trying to explain what was and what wasn't different about their research process in comparison with the standard procedures in their field" (Harding 1993, p. 51). This criticism was a form of empiricism, according to Harding, because it accepted certain basic characterizations of science that were accepted both by scientists and by the majority of philosophers of science of that time. Harding characterizes empiricism as being dedicated to the belief that science, or at least good science, is value-free. The social identity and the personal values of scientists would therefore be irrelevant to science. What mattered were criteria like the success of predictions and the ability of theories to provide a relatively simple explanation for a broad and diverse range of phenomena. Harding further characterizes spontaneous feminist empiricism as being concerned with case studies of bad science, believing that "sexism and androcentrism are social biases correctable by stricter adherence to the *existing* methodological norms of scientific inquiry" (Harding 1986, p. 24, emphasis added). In other words, the androcentric science critiqued by Fee and by Fausto-Sterling was the result of science's failure to live up to its own standards.

While Harding acknowledged the contributions to feminist aims that were made by spontaneous feminist empiricism, her own preference was for a critical approach that took greater account of the social environment within which science was practiced. As noted earlier, traditional empiricism in philosophy of science viewed the products of scientific work as being value-free. While scientists' personal interests and biases could inform the early stages of their work, the nature of the scientific method meant that these values were winnowed out of the final theory. This view was reflected in the traditional distinction between the context of discovery, which focuses on how theories are generated and the context of justification, which looks at how theories are tested. By contrast, Harding preferred a form of feminist standpoint theory, which emphasizes the pervasive influence of the social organization of scientific practice on scientific theories. Rather than seeing science as ultimately based on careful, neutral observations and logical argument, feminist standpoint theory analyzed science as a social practice in which members of social groups that have been traditionally excluded or marginalized can offer "a morally and scientifically preferable grounding for our interpretations and explanations of nature and social life" (Harding 1986, p. 26). Through political action, women and other marginalized groups can come to recognize both their distinctive social position and the advantage that that position confers. Harding contrasted this explicitly social and political approach to understanding science with that taken by

spontaneous feminist empiricism. Whereas the latter focused, not on changing science as a whole but on exposing cases of bad scientific research, standpoint theory aimed to replace androcentric methods and approaches to evidence with a more adequate account based on the standards suggested by women's experiences.

However, the distinction between spontaneous feminist empiricism and standpoint theory soon needed to be revised to reflect theoretical developments in feminist philosophy of science. At the same time that feminist standpoint theory was being developed, other feminist philosophers of science were also developing theoretically sophisticated forms of feminist empiricism that took into consideration both the political concerns of standpoint theory and the emphasis of traditional philosophy of science on standards for the adequacy of evidence.

Philosophical feminist empiricism, as exemplified in work by Helen Longino (1990, 2001) and by Lynn Hankinson Nelson (1992), differed from both traditional empiricism and spontaneous feminist empiricism in that it recognized that social values and interests play a role in science but continued to emphasize the fundamental importance of empirical criteria in choosing and testing scientific theories. It also recognizes the importance of epistemic values that are conducive to the empirical success of science, such as those discussed above (simplicity, explanatory power, etc.), though Helen Longino (1996) has developed a distinct set of feminist epistemic values that may sometimes be in tension with the more traditional list. While Harding does not emphasize this point, both Longino and Nelson also emphasize (in different ways) that knowledge must be analyzed at the level of the scientific community, rather than that of the individual scientist.

Although both philosophical feminist empiricism and feminist standpoint theory recognized the social nature of scientific knowledge and the inescapable role of values in science, some tension remained between them. Harding argued that feminist standpoint theory was the best way to challenge the traditional empiricist (and value-free) conception of science because of its more radical challenge to scientific methods and its emphasis on the influence of social location (and of the standpoint to which a social location might lead) on knowledge. She also argued that the implication of spontaneous feminist empiricism, which followed traditional empiricism, the social identity of scientific inquirers was "irrelevant to the 'goodness' of the results of research," sat uneasily with claims that a feminist approach to science would reduce instances of scientific bias. By contrast, Nelson worried that standpoint epistemologies, by positing an "epistemological chasm between feminists and nonfeminists" so that "there are things one group can know and another cannot know" (1994, p. 41), gave up any hope of finding an empirical common ground.

These disagreements are important to the history of feminist philosophy of science, but they should not be overestimated. It should be noted that in developing her defense of standpoint theory, Harding contrasts it mainly with spontaneous feminist empiricism and simply notes that "[i]t would be an interesting and valuable project to contrast in greater detail these important philosophical feminist empiricisms and/with feminist standpoint theory" (1986, p. 52); she does not undertake this latter project. More recently, Kristen Intemann (2010) and Elizabeth Potter (2006) have argued that developments over the past two decades in both

(philosophical) feminist empiricism and standpoint theory have brought the two approaches closer together. Intemann suggests that adherents to each view may have misrepresented the other, and that, in reality, “the main tenets of standpoint theory are far closer to those of feminist empiricism than proponents of either view have been willing to admit” (Intemann 2010, p. 206). She also notes that people working in the two different traditions also modified and clarified their views in response to criticism from the other side. It is perhaps best to view the two approaches as developing over time and as coming closer together as they did so. The two strands of feminist philosophy now differ mainly in emphasis, with standpoint theory focusing on the importance of social location and standpoint in shaping science and feminist empiricism emphasizing the importance of judging theories by their empirical successes, i.e., by their ability to make successful empirical predictions.

Some Central Insights of Feminist Philosophy of Sciences

Despite these differences in emphasis, there are a number of areas of agreement between feminist standpoint theory and feminist empiricism, and these can be taken to be core commitments of feminist philosophy of science.

The Role of Values in Science

As Elizabeth Potter (2006) notes, much of feminist philosophy of science starts from an examination of the relationship between science and values. In particular, feminist philosophers emphasize that science is shaped by various kinds of values (social, political, epistemic) at all stages of inquiry. Remember that the traditional empiricist approach to philosophy of science held that values were eliminated in the process of justifying theories; failure of this process meant that the resulting theory was an example of bad science.

By contrast with traditional empiricism, feminist philosophers viewed the influence of values on science as inescapable, and discussion of this problem has been one of the main areas of research in feminist philosophy of science. This meant that it was impossible to just dismiss the kind of science critiqued by the early (spontaneous) feminist empiricists as bad simply because it reflected social values. Instead, there had to be a way to differentiate between legitimate and illegitimate influences of values in science. Elizabeth Anderson (2004) has argued that values exert an illegitimate influence on scientific research when they drive the research to a predetermined conclusion. The craniologists whose work Elizabeth Fee examined looked for evidence to support their social beliefs about the relative intelligence of women and men and also thought that this evidence could be supplied by finding differences in the sexes’ skulls. Their unwillingness to give up their scientific hypothesis and the pattern of reformulating the specific changes they expected to see are evidence that their values pushed their research to the only conclusion they would accept.

This type of influence also occurs in contemporary research. For example, functional magnetic resonance imaging (fMRI) studies that look at (or for) sex/gender differences in responses to emotional stimuli generally start from the simple claim that women are more emotional than men, despite the complex and equivocal nature of much of the psychological literature on sex/gender differences in emotion, and look for differences in brain activity to explain these purported psychological differences. Despite the lack of clear evidence of differences in many of these studies, the authors always interpret their results in such a way as to support their original hypothesis of differences in brain activity (Bluhm 2012, 2013). Here again, the conclusion of the research appears to have been determined even before the study was conducted.

Although this research is clearly influenced by beliefs about gender and gender differences, its focus on differences may also be due in part to the difficulty of publishing "negative" results (i.e., results that show no difference between groups). This is another example of the way in which values determine the way science is conducted. Here, the values are not political, but they have a strong influence all the same. Moreover, even though the values themselves are not political, they have political consequences, since they lead to the conclusion that there are sex/gender differences.

A closer look at these studies shows that their very design makes it easy for researchers to find *some* sort of brain activity difference to talk about. Because the hypothesis being examined in these studies is that there are brain activity differences either "somewhere" in the brain, or in particular areas previously associated with emotion processing, any difference in observed brain activity can be taken to support the claim that differences in emotionality can be explained by differences in brain activation. This area of research has to develop more sophisticated models of emotion processing in order to provide clear evidence for the claim that women and men process emotional stimuli differently (Bluhm 2013). When hypotheses appeal only to vague activity differences, any difference can be taken in support of a sweeping claim regarding psychological differences. This does not mean that the researchers should not have conducted research on sex/gender differences in the first place, but it does mean that they should have been more open to disconfirming evidence and should have designed their studies so that their hypothesis was not confirmed by just about *any* evidence.

Though in fact, given how *much* disconfirming evidence there tends to be in sex/gender difference research, it is not unreasonable to ask why researchers are still working in this area. One possible reason is that sex/gender differences in emotion processing may have practical implications. Many of the fMRI studies I examined note in the introduction that women are much more likely than men to experience depression and suggest that this may be due in part to sex/gender differences in emotion processing. They often further suggest that a better understanding of emotion processing may lead to improved treatment of depression. Here is another example of the influence of values in science. Because scientists are encouraged – and even required by some funding agencies – to address the possible practical applications of their work, they often resort to speculation in cases where there is not direct potential

applicability. (This often occurs, as well, in work using animal models, which may indeed result in important knowledge, but which is not directly applicable to humans).

Yet, in fact, the researchers do not tend to take these issues into consideration in designing their studies, perhaps by including participants with depression, or attempt to incorporate neuroimaging studies that do look at brain activity in depression in their discussion of their results. It does not seem, then, that there is an honest attempt to address potential practical implications, so the research is clearly not motivated by these problems. Rather, as Fausto-Sterling said á propos of research on the corpus callosum, the persistence of this line of research seems to be evidence only of researchers' commitment to the belief that sex differences exist. And as with the other research I have surveyed, this belief persists despite any number of studies that fail to show clear sex/gender differences.

This brings us to a second aspect of Anderson's discussion of science and values. As noted above, Anderson suggests that values play an illegitimate role in science when they push research to a predetermined conclusion. This may be a sign that the values in question are held dogmatically (albeit possibly unconsciously) and scientists thus fail to change them in the face of evidence to the contrary. The belief (which also reflects political values regarding the existence of sex/gender differences) that "women are more emotional than men" may simply be held dogmatically; it appears to be held in a fairly unsophisticated form despite the lack of clear evidence to support it. Anderson emphasizes that just as science is shaped by values, values should be also responsive to empirical evidence. In cases where values are responsive to evidence, we can expect that scientists whose research does not show promising results will look for other avenues to investigate. This does not seem to happen with research on sex/gender differences.

Objectivity

A second area in which feminist philosophy has made important contributions is in understanding the social nature of science and its implications of this sociality for objectivity. Here again, we can start from the question of what sorts of values should play a role in science. If we take "objective" to mean value-free, or neutral, then it would appear that science can never be objective. But this conclusion is too quick. Since different individuals and different groups have different values, feminist philosophers of science have argued that science becomes more objective as *more* values are added to it. Fee's craniologists all shared the same basic (sexist) values, so their way of understanding the phenomenon they were studying went unchallenged (see also Okruhlik (1994)). Including researchers with different approaches and values can therefore help researchers to design better experiments, for example, by sharpening their hypotheses or qualifying their conclusions more carefully. For example, in the case of the fMRI research that I looked at, discussed above, the researchers may have designed better experiments if they had consulted with social psychologists, who study just how complex and nuanced emotional

experiences and expression can be and the extent to which sex/gender differences depend on environmental variables (see, e.g., Brody 1999).

The most influential discussions of objectivity in feminist philosophy are those offered by Helen Longino and by Sandra Harding. Both emphasize that incorporating diverse viewpoints into scientific research will provide a better, more objective, understanding of the phenomena being studied. Harding's "strong objectivity," which is based in standpoint theory, emphasizes the need to start from the perspectives of those who have historically been excluded from science and also the need for researchers to engage in critical examination of how their own social location influences their research (1995). Longino's critical contextual empiricism focuses on developing norms for scientific inquiry that promote objectivity by increasing the possibility that background assumptions shaping the research will be recognized and debated. These norms are: the existence of public venues for science criticism, uptake of criticism by the scientific community, shared public standards by which to evaluate scientific methods and claims, and tempered equality of participants in the scientific discourse (Longino 1990, 2001). Both theories emphasize that diversity in scientific communities is the best way to prevent the illegitimate influence of values on scientific research.

While Harding, in particular, emphasizes the importance of diversity of social groups or standpoints, it is also especially important in an area like neuroscience to incorporate diversity of *scientific* viewpoints. Relevant disciplines may include a number of subfields of psychology, sociology, linguistics, anthropology, computer science, and statistics, depending on the particular question being addressed. Because the brain is studied at a variety of biological levels and because the field of cognitive science was self-consciously founded as an interdisciplinary endeavor, there is a rich set of resources available for scientists to enable them to increase the objectivity of their research. Longino's norms may help to prevent the problem noted above, in which neuroimaging researchers did not work with social psychologists to clarify their hypotheses about sex/gender differences in emotion.

"Studying Up"

A third aspect of feminist philosophy of science, related to the issues of objectivity discussed above, is its insistence that research on social groups must start from the perspective of members of that group. The need to "study up" from the vantage point of oppressed groups is one of the key distinguishing features of standpoint theory, though feminist empiricists have also come to acknowledge it. Due to social differences between many scientists and the groups they study, scientists' understanding of the phenomena they are investigating may differ from that of the people they study, with the result that the research provides only a partial or skewed perspective on what is going on.

An example of research that was explicitly based on feminist principles is Gillian Einstein's work on the neurological effects of female genital circumcision/mutilation/cutting (FGC). Einstein was interested in the effects of this procedure on the central nervous system and on the body as a whole but also in the experiences of women who had undergone the procedure. She worked with a group

of Somali immigrants in Toronto, Ontario; these women not only were research participants but were also involved in designing the study. Upon consultation with this group, Einstein chose to use a combination of qualitative and quantitative methods, because neither on their own were able to address the range of questions she was interested in answering. As a result, she was able to investigate the physical effects of FGC on the nervous system and on the rest of the body, as well as the women's experiences of life after FGC (which were influenced by their knowledge that the practice is disapproved of in their new country).

Einstein draws on her experience planning and carrying out this research to develop a number of "guideposts" for feminist neuroscience practice, which include commitments to the kind of reflexivity demanded by strong objectivity; to considering research participants as partners; to understanding variability instead of focusing on averages; and to approaching a problem from multiple perspectives (Einstein 2012, pp. 166–168). It is important to recognize that this approach is not just a reflection of feminist political commitments but that it is also a better approach from an *empirical* perspective, since it provides a more adequate understanding of the phenomena being studied than research that does not use these guideposts.

Beyond Research on Sex and Gender

The previous section summarized some of the central insights of feminist philosophy of science and showed how they apply in both criticisms of neuroscience research and developing better research projects. This section briefly describes some ways in which these insights can apply in areas of neuroscience that are not explicitly concerned with sex/gender differences or with women's biology.

A number of debates in neuroethics might be better understood if we pay close attention to the kinds of assumptions that researchers make about the nature of the phenomenon they are studying, about the best methods and techniques to use in their research, and about how their research fits with other areas of investigation. One obvious extension of work in feminist philosophy is to take what we have learned from studying research on sex/gender differences and apply it to other areas in which two or more groups are compared, such as clinical neuroimaging, research on neurodevelopment and aging, and pre-post studies that examine the effects of learning or of other interventions. There are also other areas of "applied ethics" where feminist philosophy of science may provide useful tools, including debates over neurocognitive enhancement, lie detection, the use of brain-computer interfaces, and the use of pharmaceutical interventions to alter memory (Hurley 2010). At a more abstract level, neuroethics also investigates the effects of research in neuroscience on questions that have important implications for our self-understanding, such as questions about free will and moral responsibility, personal identity, and the possibility of "mind reading" using neuroimaging technologies. All of these are complex concepts, and feminist philosophy of science may be helpful in clarifying the ways in which scientists' preconceptions about these questions shape their research.

Finally, any research that explicitly deals with social questions or focuses on particular groups may benefit from feminist analysis. The idea of "studying up" means that the viewpoint of participants should be included in designing and conducting neuroscience research, which may mean including questions that are important to patients in clinical studies or collecting phenomenological data in order to learn how study participants interpret the task (Womack and Mulvaney-Day 2012; Gallagher 2012). And, finally, a feminist approach will also be useful for research in social cognitive affective neuroscience (SCAN), which needs to frame questions that can have clear political implications in a way that is methodologically tractable but that avoids oversimplifying complex issues and questions. Specific topics here include the nature of social stereotypes, the evolution of emotion and of moral reasoning, and the relationship between psychopathy and theory of mind, as well as consideration of sex/gender differences (Grossi and Fine 2012).

Conclusion

Feminists have had much to say about neuroscience research, and feminist philosophy of science has drawn, in part, on issues arising from neuroscience in developing its particular approach. Starting from the insight that science cannot – and should not – be value-free, feminist philosophy of science has developed a number of ways to understand the influence of moral, political, and methodological values on the way science is conducted. This paper has argued that these tools can be very useful in understanding questions in neuroethics.

Cross-References

- ▶ Experimentation in Cognitive Neuroscience and Cognitive Neurobiology
- ▶ Feminist Neuroethics: Introduction
- ▶ Historical and Ethical Perspectives of Modern Neuroimaging
- ▶ Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences
- ▶ Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform

References

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia, 1*(19), 1–24.

Bell, A. D., & Variend, S. (1985). Failure to demonstrate sexual dimorphism of the corpus callosum in childhood. *Journal of Anatomy, 143*, 143–147.

Birke, L. (1986). *Women, feminism and biology: The Feminist challenge*. Brighton: Wheatsheaf Books.

Bleier, R. (1984). *Science and gender: A critique of biology and its theories on women*. Oxford: Pergamon Press.

Bluhm, R. (2012). Self-fulfilling prophecies: The influence of gender stereotypes on functional neuroimaging research on emotion. *Hypatia*. doi:10.1111/j.1527-2001.2012.01311.x. Published online ahead of print.

Bluhm, R. (2013). New research, old problems: Methodological and ethical issues in fMRI research examining sex/gender differences in emotion processing. *Neuroethics, 6*(2), 319–330.

Brody, L. (1999). *Gender, emotion, and the family*. Cambridge: Harvard University Press.

DesAutels, P. (2010). Sex differences and neuroethics. *Philosophical Psychology, 23*(1), 95–111.

Einstein, G. (2012). Situated neuroscience: Exploring biologies of diversity. In R. Bluhm, A. J. Jacobson, & H. L. Maibom (Eds.), *Neurofeminism: Issues at the intersection of feminist theory and cognitive science* (pp. 145–174). Basingstoke: Palgrave Macmillan.

Fausto-Sterling, A. (2000). *Sexing the body: Gender politics and the construction of sexuality*. New York: Basic Books.

Fee, E. (1979). Nineteenth century craniology: The study of the female skull. *Bulletin of the History of Medicine, 53*, 415–433.

Gallagher, S. (2012). Taking stock of phenomenology futures. *The Southern Journal of Philosophy, 52*(2), 304–318.

Grossi, G., & Fine, C. (2012). The role of fetal testosterone in the development of the "essential differences" between the sexes: Some essential issues. In R. Bluhm, A. J. Jacobson, & H. L. Maibom (Eds.), *Neurofeminism: Issues at the intersection of feminist theory and cognitive science* (pp. 73–104). Basingstoke: Palgrave Macmillan.

Harding, S. (1986). *The science question in feminism*. Ithaca: Cornell University Press.

Harding, S. (1993). Rethinking standpoint epistemology: What is strong objectivity. In L. Alcoff & E. Potter (Eds.), *Feminist epistemologies* (pp. 49–82). New York: Routledge.

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese, 104*(3), 331–349.

Hurley, E. A. (2010). Pharmacotherapy to blunt memories of sexual violence: What's a feminist to think? *Hypatia, 25*(3), 527–552.

Intemann, K. (2010). Standpoint empiricism: Rethinking the terrain in feminist philosophy of science. In P. D. Magnus & J. Busch (Eds.), *New waves in philosophy of science* (pp. 198–225). Hampshire: Palgrave MacMillan.

Longino, H. E. (1990). *Science as social knowledge*. Princeton: Princeton University Press.

Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 39–58). Dordrecht: Kluwer.

Longino, H. E. (2001). *The fate of knowledge*. Princeton: Princeton University Press.

Nelson, L. H. (1992). *Who knows: From quine to a feminist empiricism*. Philadelphia: Temple University Press.

Nikoleyczik, K. (2012). Towards diffractive transdisciplinarity: Integrating gender knowledge into the practice of neuroscientific research. *Neuroethics, 5*(3), 231–245.

Okruhlik, K. (1994). Gender and the biological sciences. *Canadian Journal of Philosophy, 20*(Suppl.), 21–42.

Potter, E. (2006). *Feminism and philosophy of science: An introduction*. New York: Routledge.

Roy, D. (2012). Neuroethics, gender and the response to difference. *Neuroethics, 5*(3), 217–230.

Sayers, J. (1982). *Biological politics: Feminist and anti-feminist perspectives*. New York: Tavistock Publications.

Tuana, N. (1993). *The less noble sex: Scientific, religious, and philosophical conceptions of women's nature*. Bloomington: Indianan University Press.

Womack, C., & Mulvaney-Day, N. (2012). Feminist bioethics meets experimental philosophy: Benefits of embracing the qualitative and experimental. *International Journal for Feminist Approaches to Bioethics, 5*(1), 113–132.

Feminist Ethics and Neuroethics

89

Peggy DesAutels

Contents

Abstract

Do men and women have distinct innate moral psychologies? Both Aristotle and Kant assumed so and viewed women as psychologically incapable of being fully moral. Conversely, John Stuart Mill and Carol Gilligan argued that women are equally as capable of morality as men, but also point out that women are nonetheless psychologically influenced by gendered roles and expectations. It is perhaps the most significant insight in feminist moral psychology that the distinction between the psychologies of oppressors and those they oppress is a much more significant psychological distinction to make and analyze than is the distinction of the moral psychologies of biologically sexed women and men. Moreover, the moral psychologies of oppressors and the oppressed are damaged in very specific ways. The psychologies of domestically abusive men are a case in point. Male abusiveness appears to be trait-like, with testosterone levels making only a minor contribution to the maintaining of this trait. The degree to

P. DesAutels
Department of Philosophy, University of Dayton, Dayton, OH, USA
e-mail: peggy.desautels@gmail.com; Peggy.DesAutels@notes.udayton.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_155,

which a society is patriarchal and the role models to which men have been exposed play much larger roles in the development of the brains of abusive men.

Introduction

Neuroethics, as a field, is broad and not yet fully defined. There is, however, one clearly emerging branch of neuroethics devoted to the neuroscience of moral judgments and behaviors. The focus in this chapter is on a feminist approach to this particular branch of neuroethics. More specifically, this chapter examines the following: (1) if and to what degree there are sex and gender differences in moral judgments, behaviors, and traits and (2) if there are apparent differences, what are the most likely causes? The case is made that most differences are attributable not to sex-based brain differences between men and women, but to distinct and persistent types of psychologies found in oppressors and the oppressed.

It has long been assumed that women and men have different types of moral (and immoral) psychologies. Women have been assumed to be, in general, the more caring, nurturing, and empathetic sex; and men have been assumed, in general, to be the more rational, in-control, and duty-oriented sex. Conversely, women are stereotypically assumed to be the more manipulative, petty, and contriving sex; and men are stereotypically assumed to be the more violent, abusive, and aggressive sex. As a result of these widely held assumptions, there are sex-differentiated behavioral norms and virtues. Women are expected to cultivate such "feminine" virtues as compassion, emotional supportiveness, and kindness; and men are expected to cultivate such "male" virtues as strength-of-will, leadership, and courage. But are these supposed sex-differentiated moral psychologies and accompanying sex-differentiated virtues grounded on any brain-based, innate sex differences?

To answer this question, this chapter begins with a very brief history of gender-related psychological assumptions and norms found in the ethics of Aristotle (384–322 B.C.) and Kant (1724–1804). Both assumed that women are incapable of being fully moral and thus need the moral guidance of men. The chapter includes an even briefer history of feminist responses to the view that women are incapable of being fully moral, including the feminist responses of John Stuart Mill (1806–1873) and Carol Gilligan (1936–). Next there is a summary of perhaps the most significant development in feminist moral psychology: the insight that the distinction between the psychologies of oppressors and those they oppress is a much more significant psychological distinction to make and analyze than is the distinction between the moral psychologies of biologically sexed women and men. Moreover, the moral psychologies of oppressors and the oppressed are damaged in very specific ways. The chapter concludes with an analysis of the psychologies of abusive men to determine the degree to which male abusiveness is trait-like and the respective roles that testosterone and patriarchal societies play in the development of the brains of abusive men.

A Brief History of Gender and Ethics

Aristotle's virtue approach to ethics has been influential throughout Western history and remains influential in ethical theory today. Unfortunately, Aristotle's negative views on women have also been influential throughout much of Western history. Although Plato and a few other early male philosophers held more enlightened views on women, Aristotle's views were most influential on the views of Aquinas (1225–1274) and thus on the Catholic Church and Western society more broadly. Women, on Aristotle's view, are incomplete males who lack the ability to be fully rational and moral. In comparison to men, women are "more void of shame and self-respect, more false of speech, more deceptive, and of more retentive memory" (HA IX). In addition, unlike men, women are unable to exercise full control of reason over their appetites (NE, VII, 7–8; Pol, I, 13 (1259b 35–38)). Thus, according to Aristotle, it was morally necessary for women to be obedient to men, and it was impossible for women to engage in equal political or personal relationships with men. Aristotle viewed the marital relationship between a man and a woman as an unequal friendship in which the wife is expected to love the husband more than the husband loves the wife, and the wife is expected to recognize more honor and authority in her husband than he recognizes in her. Maryanne Cline Horowitz writes, "Aristotle's biological and psychological ideas about women parallel his political and ethical ideas about women. Together, these ideas are circular, self-supporting, and antifeminist to the core. On the one hand, woman's alleged inadequacies of body and mind backed up his general dictum that men naturally rule over women. On the other hand, Aristotle's unwillingness to view women as potential voting and ruling citizens and his unwillingness to recognize alternatives to the political and ethical hierarchy of husband over wife limited his viewpoint on female capacities" (Horowitz 1976, p. 212).

Kant's moral philosophy emphasizes the importance of rationality (detached from emotion and inclination) in moral deliberations. Like Aristotle and reflective of the societal norms of his times, Kant defended the views that husbands should be their wives' masters and that women should be denied citizenship. Also, like Aristotle, Kant viewed women as more likely than men to be swayed by emotions and passions and thus less likely than men to be motivated by rationality. Because of their psychological natures, women need the moral, legal, and political guidance of men. Unsurprisingly, he prescribed sex-differentiated virtues. Women were expected to emphasize the cultivation of the feminine virtues of charm, modesty, and subservience – virtues that benefitted men. Lawrence Blum, in a feminist analysis of both Kant's and Hegel's ethics, points out, in addition, that these philosophers' versions of moral rationalism are male-centered in two ways: "First, [moral rationalism] places at the center of its scheme of virtues qualities of character—rationality, strength of will, adherence to universal principle and duty—which can be seen . . . as 'male.' Correlatively it fails to provide a framework for expressing the significance of an important range of human virtues, which can be seen as 'female.' Such virtues are sympathy, compassion, human

concern, kindness, and emotional supportiveness. Second, it is a philosophy which reflects a male-dominated society, and implicitly sanctions male superiority" (Blum 1982, 294). For more on Kant's views on women and ethics, see Cash 2002.

Feminist Responses

John Stuart Mill was one of the first significant figures in the modern history of ethics to challenge the view that women were less capable of morality than men and to point out that both women's and men's moral psychologies are damaged by patriarchal societies. He writes that in order for women to be the obedient and "willing slaves" of men, men need to enslave women's minds by educating them to be subservient. He writes, "All women are brought up from the very earliest years in the belief that their ideal of character is the very opposite to that of men; not self-will, and government by self-control, but submission, and yielding to the control of others. All the moralities tell them that it is the duty of women, and all the current sentimentalities that it is their nature, to live for others; to make complete abnegation of themselves, and to have no life but in their affections" (Mill 1869, par. 25). When men rule in society, it is important that women's psychologies accept this and willingly accommodate a man-as-master and woman-as-slave arrangement of social life.

According to Mill, when men rule over women, it is not only women's moral psychologies that are damaged, but men's moral characters as well. He notes that "All the selfish propensities, the self-worship, the unjust self-preference, which exist among mankind, have their source and root in, and derive their principal nourishment from, the present constitution of the relation between men and women. Think what it is to a boy, to grow up to manhood in the belief that without any merit or any exertion of his own, though he may be the most frivolous and empty or the most ignorant and stolid of mankind, by the mere fact of being born a male he is by right the superior of all and every one of an entire half of the human race. . .What must be the effect on his character, of this lesson?" (Mill 1869, par. 33). Mill points out that it is impossible to know what men's and women's character traits would be if women were viewed as fully equal to men both in politics and in households, since no society has ever been set up that way.

Mill's views are a good historical lead-in to more recent developments in feminist ethics. Many of these recent developments grow out of the groundbreaking work of Carol Gilligan, a psychologist, who published *In a Different Voice: Psychological Theory and Women's Development* in 1982. Gilligan studied under and critiqued Lawrence Kohlberg (1927–1987), who proposed six main stages of moral development (see Kohlberg et al. 1983). Kohlberg's highest proposed stage was Kantian and involved making moral judgments by applying universal principles to moral situations. In developing his theory, Kohlberg studied young boys exclusively. When Gilligan expanded Kohlberg's project to include girls, it appeared that girls were, in general, less morally developed than boys, since they were more likely to focus on caring relationships than abstract justice when

reasoning morally, which would put them only at Stage Three on Kohlberg's scale of moral reasoning (Gilligan 1982). Gilligan later proposed that there are two incompatible moral perspectives – the justice perspective and the care perspective – neither of which is "higher" than the other on a moral scale and that both men and women are capable of shifting between these two perspectives (Gilligan 1987). Some feminist philosophers drew on Gilligan's work to propose ethical theories that emphasize and value the role of caring relationships, which are designated "the ethic of care." Other feminists objected that caring work is denigrated and undervalued in ethical theory because it has been tied to women and to other oppressed groups, and so eschewed the ethic of care, or at least urged caution that it not be gendered. The original hypothesis that women tended to be more caring than men was replaced with the hypothesis that members of oppressed groups, including women, are expected to do society's care work, which might make it more likely that they develop the necessary psychologies to do this. However, there is no evidence that women are naturally more caring than men, nor is it indisputable that women acquire traits associated with care if they engage in the caring work that is expected of them.

In 1990, Sandra Lee Bartky analyzed the influence of patriarchy on women's psychologies – picking up where Mill left off. In her book, *Femininity and Domination: Studies in the Phenomenology of Oppression*, she examines the ways that women internalize oppression. She writes, "I have been interested from the first in the nature of that 'femininity' that disempowers us even while it seduces us; I want to understand how the values of a system that oppresses us are able to take up residence inside our minds" (Bartky 1990, p. 2). She has taken up Mill's project using a wider range of theoretical tools, including linguistics, Marxism, empirical social science, and psychoanalysis. Unlike Mill, she also provides ways that women can resist the enslaving of their minds through feminist consciousness raising.

Clearly, what is provided above is an abbreviated and oversimplified history of gender and ethics. There have been many other developments and significant insights in both feminist ethics and feminist moral psychology (see, e.g., *Moral Psychology: Feminist Ethics and Social Theory* edited by Peggy DesAutels and Walker 2004). Nonetheless, it should be quite evident that from the very beginning (Aristotle), ethics has been developed with the assumptions that men should rule women and women should be obedient and pliable. Norms of femininity even today expect women to be first and foremost attractive to men and supportive of men. Conversely, norms of masculinity require that men contain their emotions, refrain from caring sentimentality, and stay in control of themselves and those around them.

The Brains of Oppressors and the Oppressed

The psychologies of women and men develop within and are influenced by patriarchal social norms and structures. Allen Johnson summarizes the key features of a patriarchal society: "A society is patriarchal to the degree that it promotes male

privilege by being *male dominated*, *male identified*, and *male centered*. It is also organized around an obsession with control and involves as one of its key aspects the oppression of women" (Johnson 2005, p. 5). Contemporary societies, including such "developed" societies as that found in the United States today, remain patriarchies. Thus, even today and throughout the world, women's psychologies are, at least to some degree, the psychologies of the oppressed, and men's psychologies are, at least to some degree, the psychologies of oppressors. If this is the case, where does neuroscience come in? So far, many of the neuroscientific insights on moral cognition have come from fMRI studies while subjects think through moral dilemmas and make moral judgments. Of course it is impossible to conduct fMRI studies on subjects as they go about their daily lives and engage in moral and immoral behaviors. And it is also impossible to conduct fMRI studies on the cognitive processes involved in internalizing norms of femininity and masculinity. In other words, neuroscientists are going to find great difficulties in examining and understanding the brains of oppressors and the oppressed. And even if this becomes more possible through better experimental designs, there will remain the difficulty of separating nature from nurture in the development of sex-differentiated behaviors and character traits.

Most feminists would agree that the brains of women and men are biologically sexed to at least some minimal degree, but they would emphasize in addition that men's and women's brains are embedded in particular social structures with learned patterns of behavior that contribute to how these brains are organized and shaped. Perhaps the most obvious biological sex difference with the potential to differentially affect men's and women's brains is the level and types of hormones that affect brain development and organization. There is no denying that starting in the womb, males and females have differing hormones and differing hormone levels. And there is no denying that hormones play a role in brain development, organization, and functioning. But it remains unclear exactly what the cognitive effects of these hormonal differences really are. For example, Rebecca Jordan-Young's recent book, *Brain Storm: The Flaws in the Science of Sex Differences* (2010), convincingly shows that hormone-based "brain organization theory" is not as scientifically grounded as both neuroscientists and the popular media would lead us to believe. Nonetheless, some progress is being made in determining the respective roles that hormones, situational factors, and cultural norms and practices play in shaping cognition and behavior. As a result, perhaps we can make progress in answering the following sorts of questions: Do differences in men's and women's hormone levels affect their moral judgments and behaviors? Do differing levels and types of hormones absolve men and women in differing ways and degrees, of moral responsibility for their actions? What role does testosterone play in aggression and thus in the committing of violent crimes? Are more men than women in prison because their "moral brains" differ in fundamental ways?

Recently, a number of ethicists have rejected abstract and ideal theorizing in favor of more empirically informed approaches to explaining and prescribing moral cognition and behavior. Of special interest to some of them (e.g., John Doris and

Gilbert Harmann) is the degree to which personality traits or dispositions are the determinants of at least some moral behavior. Doris, for example, appeals to a number of psychological studies including the Good Samaritan experiments, the Milgram experiments, and Holocaust research, to argue that situations are much more powerful determinants of behavioral regularities than are supposed character traits (Doris 2002, p. 26). Doris focuses primarily on the ways that situations override supposed virtuous traits and seldom discusses vicious traits. Although Doris is right to say that situations play a significant role in determining moral behavior, there is also good reason to argue (contra Doris) that certain vicious patterns of behavior are tied to vicious character traits for two main reasons: (1) vicious behavior is often perceived by the vicious agent to be in that agent's best interest, and thus the agent benefits from, values, and consciously perpetuates certain vicious traits, and (2) in the case of male domestic abusers, a persistent and pervasive situation (i.e., patriarchal culture) combined with certain personal histories (i.e., exposure to abusive behaviors when growing up) combine to produce the rigorous vicious character trait of abusiveness to women that is predictive of patterns of vicious abusive behaviors. Psychological studies of male domestic abusers show that a belief in male entitlement combined with early exposures to males dominating and abusing women contributes to and perpetuates vicious character traits in certain men that result in their exhibiting regular patterns of abusive behaviors in domestic settings.

The Psychologies and Traits of Male Domestic Abusers

It is helpful at this point to provide some statistics for the United States published by the US Department of Justice:

- The number of victims of intimate partner crimes in the United States was approximately 907,000 in 2010.
- From 1994 to 2010, about four in five victims of intimate partner violence were female.
- Most female victims of intimate partner violence were previously victimized by the same offender, including 77 % of females ages 18–24, 76 % of females ages 25–34, and 81 % of females ages 35–49 (Catalano 2012).

As these statistics show, violent domestic abusers are almost exclusively male, their victims are almost exclusively female, and abusers tend to be repeat offenders. Why is that? And just how likely is it that a particular domestic abuser will continue to abuse or even escalate the abuse of his domestic partners? Knowing the answers to these questions could be lifesaving information both for the surviving victims of abuse and for the new partners of past abusers. In other words, it could be lifesaving to know whether or not men with a history of domestically abusive behavior have an abusive vicious character trait and as a result can be expected to continue to abuse their domestic partners. Thus, the answer to Doris's more general question about the degree to which behavioral regularities can be attributed to character traits

becomes potentially an answer with life-and-death consequences in the case of violent abusive behavior. Too often, women think that abusers can and will change with changed circumstances.

To address the degree to which abusive behaviors in domestic settings are tied to abusive character traits, and could be viewed as a type of Aristotelian vice, it is worth starting with Aristotle's discussions of virtues and vices. Although Aristotle's views on women are both unfounded and morally problematic, his virtue approach to ethics grounds and is relevant to contemporary discussions of virtues and vices. It is also worth looking at the latest research on brain-based causes of violent and aggressive behaviors and on the psychologies of heterosexual male abusers. The most likely conclusion is that in most cases, those men who are domestically abusive do have trait-like tendencies to be abusive and thus that they have vicious characters in an Aristotelian sense. However, these abusive character traits are not tied primarily to male androgens. Rather, they are formed within patriarchal social structures and maintained in individual males primarily by habituation.

For Aristotle, both vicious and virtuous character traits are acquired through habits often from our youth. In *Nicomachean Ethics*, Aristotle states, "A state [of character] arises from [the repetition of] similar activities. . .. It is not unimportant, then, to acquire one sort of habit or another, right from our youth; rather, it is very important, indeed all-important" (Aristotle NE, 1103b23–25). To develop a particular virtuous character trait, the goal is to find the mean between the excess and deficiency tied to the correlative vices associated with that trait. Aristotle might argue that domestically abusive men (hereafter referred to as male abusers) are vicious because they fail to find the mean tied to anger, mildness. They suffer from some combination of the excesses of anger he refers to as "irascibility," "bitterness," and "irritability." He states, "The person who is angry at the right things and towards the right people, and also in the right way, at the right time and for the right length of time, is praised" (Aristotle NE, 1125b32–33). Aristotle elaborates on the different types of people who have excess anger, "Irascible people get angry quickly, towards the wrong people, at the wrong times and more than is right; but they stop soon, and this is their best feature" (Aristotle NE, 1126a13–17). But, he notes, "Bitter people, however, are hard to reconcile, and stay angry for a long time, since they contain their emotion. It stops, however, when they pay back the offence; for the exaction of the penalty produces pleasure in place of pain, and so puts a stop to the anger. . .. This sort of person is most troublesome to himself and his closest friends. . .. The people we call irritable are those who are irritated by the wrong things, more severely and for longer than is right and are not reconciled until [the offender has suffered] a penalty and corrective treatment" (Aristotle NE, 1126a20–28). As can be seen later in this chapter, many male abusers probably best fit the "bitter" category although their vice is compounded by a tendency to exhibit controlling and dominating behaviors towards their domestic partners as well as expressions of contempt. Even though Aristotle makes it clear that men should rule women, he would not agree that men should rule women abusively, and he advocates for their having a type of friendship. He thinks that men and women's

roles in the household should be harmonious and bring pleasure to all. He notes, "Human beings. . .share a household not only for child-bearing, but also for the benefits in their life. For from the start their functions are divided, with different ones for the man and the woman; hence each supplies the other's needs by contributing a special function to the common good. Hence their friendship seems to include both utility and pleasure" (Aristotle NE, 1162a21–24). When Aristotle's view that men should rule women is updated to the view that domestic partners should have equal status, there is much that can be learned from Aristotle's extended discussion of friendship. Certainly, abusive men are not virtuous friends of their wives, girlfriends, or lovers, and Aristotle would most likely judge abusive men to have vicious character traits.

Building on Aristotle's views on virtue and character traits, it is now possible to turn to more contemporary discussions of a virtue approach to ethics. In particular, it is relevant to turn to some of Doris's views found in his book, *Lack of Character: Personality & Moral Behavior* (2002). According to Doris, "virtues are supposed to be *robust* traits; if a person has a robust trait, they can be confidently expected to display trait-relevant behavior across a wide variety of trait-relevant situations, even where some or all of these situations are not optimally conducive to such behavior" (Doris 2002, p. 18). So, by Doris's account, in order for male abusers to have vicious character traits, it is necessary to determine if male abusers "display trait-relevant behavior across a wide variety of trait-relevant situations." This could be a tricky determination. One feature of most male abusers is that this abusive trait only shows up in domestic settings. Work colleagues and extended family members often have no clue that their colleague or relative is abusive, since this behavior is not exhibited outside of the home. So, in one sense, male abusiveness is not a global trait. On the other hand, it is not, as Doris would put it, an "extremely grained" merely "temporally stable" local trait. Abusiveness does consistently and persistently occur in "trait-relevant situations." It appears that the trait-relevant situation for a domestic abuser just is the situation of being in the privacy of his own home in the company of his female partner. A significant portion of his life takes place in this situation. If a man exhibits persistent patterns of abuse with a female partner over time, or if the abuser changes his female partner, and the abuse continues with the new partner, this counts as being robust (in the sense of constancy) in "trait-relevant situations," and he should therefore be seen to have the vicious character trait of being a domestic abuser. This trait also has what Doris terms, "an evaluative dimension," when the male abuser values being abusive. It is shown in what follows that he does indeed value being abusive because abusiveness as a trait is in his self-interest to maintain. Doris believes that almost all behavioral regularities are explained by situational regularities (Doris 2002, p. 26). The perspective taken in this chapter, on the other hand, is one in which male abusiveness is viewed as a "robust dispositional structure" developed from youth, maintained by habit, valued for its self-interested ends, and predictive of future patterns of behavior regardless of the specific domestic situations in which or the specific female partners with which male abusers find themselves.

Neuroscience and Male Abusers

What do contemporary psychologists and neuroscientists have to say about male abusers? Do domestic abusers merely exhibit what Aristotle terms "an excess of anger"? If so, would "anger management" classes best address and treat their abusiveness? Are there other aspects to their psychologies that contribute to their being abusive? Why are abusers almost exclusively male and their victims almost exclusively female? The answers to these and related questions are difficult to tease out of the psychological and neuroscientific literature. The answer to why abusers tend to be male might lie in a closer look at male versus female brains. Perhaps circulating levels of testosterone (really androgens more generally) causally correlate with levels of aggression, or early exposure to testosterone explains how boys' brains become permanently wired for more aggression than girls' brains do. And perhaps high levels of aggression causally correlate with high levels of abusiveness. Then, instead of curtailing the expression of felt aggression in some way, all we would need to do to prevent domestic abuse is lower the testosterone of abusers (or raise their levels of estrogen).

Unfortunately (or fortunately?), this simple causal picture falls apart rather quickly. For one thing, male abusers tend not to be overly aggressive or angry in nondomestic settings. This would mean it is not a simple brain chemistry problem, since if it were, the aggressive and angry behaviors would show up more globally. For another thing, even if abusiveness were tied to being overly aggressive, neuroscientists and psychologists have multiple and often-conflicting measurements for aggression and publish often-conflicting studies on the relationship between high levels of aggression and high levels of testosterone. In their article on testosterone and social behavior, Booth et al. (2006) summarize the findings as follows:

> "Testosterone poisoning," now part of the language is a popular explanation for excessive "manly" behaviors such as boasting, violence and pugnaciousness. . .. In fact, there is little empirical support for these popular assertions. We cannot say that they are all false because research literature is not conclusive. But it is already clear that there is no simple one-to-one relationship between testosterone and machismo or aggressiveness or sexuality. It seems wiser to view testosterone as one component in a confluence of interacting physiological, psychological and social influences that affect behavior. (Booth et al. 2006, p. 167)

In further addressing the role that social influences play, they state:

> We do not assume that testosterone is a mechanism in and of itself that causes or creates behavior. Instead, we assume that testosterone increases the likelihood that certain behaviors will be expressed, if the propensity for that behavior already exists, and the expression of that behavior is consistent with social contextual demands. Characteristics of the social landscape permit, stimulate, suppress or set the stage for the expression of specific testosterone-behavior relationships. (Booth et al. 2006, p. 170)

For a more detailed and up-to-date analysis of the role that testosterone does and does not play in male aggression more generally, see Jordan-Young (2010).

So, based on this and a number of other related articles, it appears that testosterone is one factor among many in a complex confluence of factors (both brain based and culturally based) that underlie behavior. Most significantly, testosterone

levels alone do not directly cause or create behaviors. One way to think of the causal role of hormones in behaviors is that both men and women are in “hormonal situations” as well as social situations and that hormonal situations can and do influence patterns of behavior to at least some if not a very limited extent. It is worth extending Doris’s view of situations to include brain chemicals as one of many confluent causal factors affecting an agent’s cognition and behaviors. These chemicals would include not just hormones but any naturally occurring neurochemicals that affect mood, cognition, and behavior such as oxytocin, dopamine, and endogenous opiates. For more on the variety of neurobiological contributors to social and moral behaviors, see Suhler and Churchland (2011).

Male Abusers and Entitled Control

If an abuser’s “testosterone situation” is only of slight to no relevance to his abusiveness, what causes and maintains male abusive traits? Can male abusers be changed? There are a number of books and articles in the psychology literature attempting to answer these questions often with the motivation of providing psychologically accurate explanations for and guidance to abused women. Lundy Bancroft’s book, *Why Does He Do That? Inside the Minds of Angry and Controlling Men* (2002), is especially insightful. Bancroft has worked as a “counselor, evaluator, and investigator” for over 2,000 cases of what he terms “angry and controlling men” (Bancroft 2002, p. xvii). Although there is a range of patterns of abusive behavior and types of abusers, male abusers have much in common. Bancroft and others who analyze abusive behavior stress that abusers are mostly focused on control. Here’s Bancroft’s summary of the general pattern of behavior found in almost all abusers:

> Periods of relative calm are followed by a few days or weeks in which the abuser becomes increasingly irritable. As his tension builds, it takes less and less to set him off on a tirade of insults. His excuses for not carrying his weight mount up, and his criticism and displeasure seem constant. . .. One day he finally hits his limit, often over the most trivial issue, and he bursts out with screaming, disgusting and hurtful put-downs, or frightening aggression. If he is a violent abuser, he turns himself loose to knock over chairs, hurl objects, punch holes in walls, or assault his partner directly, leaving her scared to death. After he has purged himself, he typically acts ashamed or regretful about his cruelty or violence, at least in the early years of a relationship. . .. You then begin to see the signs of his next slow slide back into abuse, and your anxiety and confusion rise again.. . .. The abusive man tends to mentally collect resentments toward you until he feels that you deserve a punishment. After he blows, the abuser absolves himself of guilt by thinking of himself as having lost control, the victim of his partner’s provocations or his own intolerable pain. . .. “There is only so much a man can take.” (Bancroft 2002, pp. 147–148)

Notice how much of this description fits with Aristotle’s description of a bitter man: “Bitter people, however, are hard to reconcile, and stay angry for a long time, since they contain their emotion. The anger stops, however, when they pay back the offence; for the exaction of the penalty produces pleasure in place of pain, and so puts a stop to the anger. . .. This sort of person is most troublesome to himself and

his closest friends" (Aristotle, NE, 1126a20–27). "Troublesome" is one way to put it. "Terrifying and life-threatening" is another. Notice, too, that Aristotle does not connect this behavior with abusive or controlling behavior. Contemporary discussions of abusive men stress their need to dominate and control women.

Bancroft disabuses his reader of a number of myths about such men. One of these myths is that they have violent, explosive personalities. He points out that most abusive men are perfectly calm and rational in nondomestic settings (Bancroft 2002, p. 32). He also points out that abusers do not lose control; rather they consciously give themselves permission to embark on abusive rampages, both physical and verbal. For example, when police show up at the door, they are perfectly capable of immediately reverting to "in-control" behavior and then picking up where they left off once the police leave. Bancroft stresses, "an abuser almost never does anything that he himself considers morally unacceptable. He may hide what he does because he thinks other people would disagree with it, but he feels justified inside.... In short, *an abuser's core problem is that he has a distorted sense of right and wrong*.... the abuser's problem lies above all in his belief that controlling or abusing his female partner is *justifiable*" (Bancroft 2002, pp. 33–35). It could be added that the belief that men are justified in abusing women is closely related to gender role stereotypes that are promulgated in society, including that a man should be the king of his castle with entitled control over those who live in his castle. For a less clinical and more scholarly discussion of male abusers as coercive controllers of women, see Evan Stark's *Coercive Control: How Men Entrap Women in Personal Life* (2009).

One reason abusiveness is trait-like in its persistence is because abusers benefit in a number of ways from being abusive. As Bancroft points out, they become attached to privileges and benefits that come from getting their way, controlling their partners and making their own needs top priority. It is extremely difficult to treat male abusers. Bancroft asserts, "*...If we want abusers to change, we will have to require them to give up the luxury of exploitation*" (Bancroft 2002, pp. 151–157). He also points out that, "they have habits of mind that make it difficult for them to imagine being in a respectful and equal relationship with a woman" (Bancroft 2002, pp. xx–xxi). From their perspectives, they *deserve* to rule, to apply different standards of behavior to themselves than they do to women, and to treat women with contempt. This perspective is learned from and reinforced by abusive fathers or stepfathers and from a patriarchal society more generally. One of Aristotle's main points is that a man's ability to become virtuous depends on his having good role models.

Conclusion

In conclusion, it appears that the majority of male abusers have robust abusive traits with no incentive to modify these traits and much to gain by maintaining these traits. Testosterone may be a minor contributor, but it is not the main culprit. Rather, most abusers develop ingrained habits of abuse at an early age modeling their attitudes and behaviors on those of close family members. In addition, these abusive traits are developed and maintained in an all-pervasive situation – that of a patriarchal culture

in which physically stronger men have power and rule over women in the media, pornography, politics, and religion. More generally, it would seem that vicious behaviors that significantly advantage agents are more likely to become robust traits than are virtuous behaviors. Clearly, abusive vicious traits advantage those in power and help to maintain that power over those who are subject to this power.

Situationists (i.e., those who attribute behavior more to situations than to supposed character traits) such as Doris have much less of a case than they think they do. They, like most virtue theorists, have been looking at the wrong kinds of traits. It is relatively easy to find situations that make a "good" person behave badly. It is much more difficult to find situations that make a powerful "bad" person behave well. Traditionally, virtue theorists have focused on the virtues, but as Robin Dillon points out, virtue ethicists including situationists need to develop theories on *both* virtues and vices. Dillon writes, "We need to understand not only what kinds of persons to be but also what kinds not to be, not only what traits foster resistance to oppression and emancipation but also what traits support domination and thwart emancipation, and what cost to character struggles for liberation and struggles to maintain domination might exact" (Dillon 2012, p. 83). It could simply be added that when abusive and controlling traits show up in members of oppressive groups (e.g., particular men in patriarchal societies), they are likely to be more persistent and thoroughly trait-like than are their virtues.

This chapter ends with a proposed agenda for those wishing to further pursue a feminist approach to the neuroscience of ethics. Future efforts could be directed towards the following: (1) critically examining ways that neuroscientific research on moral psychologies is embedded within and contributes to gender-based social biases and injustices, including oppressive norms tied to "feminine" and "masculine" brains, and (2) developing theories of "moral brains" that are informed both by neuroscientific findings (including findings of sex-based differences) and by ethical theories (including feminist theories). Socially and culturally attuned feminist neuroscience will help to ameliorate tendencies to oversimplify the causal roles of biology (e.g., hormones) in sex-differentiated patterns of behavior. This chapter provides just one example of how it is possible to use a feminist approach to analyze sex-based differences in moral psychologies and behaviors, but there are many other examples worth analyzing.

Acknowledgement Thanks to Anita M. Superson, Robyn Blum, Lane DesAutels, and Robert C. Richardson for providing valuable feedback on earlier drafts of this chapter.

Cross-References

► Brain Research on Morality and Cognition
► Moral Cognition: Introduction
► Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences
► No Excuses: Performance Mistakes in Morality

▶ Prediction of Antisocial Behavior
▶ Psychology and the Aims of Normative Ethics

References

Aristotle. (1908a). Eudemian ethics (EE). In W. D. Ross (Ed.), *The works of Aristotle*. Oxford: Clarendon.

Aristotle. (1908b). Historia animalium (HA). In W. D. Ross (Ed.), *The works of Aristotle*. Oxford: Clarendon.

Aristotle. (1985). *Nicomachean ethics (NE)* (trans: Irwin, T.). Indianapolis: Hackett.

Aristotle. (1948). *The politics of aristotle (Pol*) (trans: Barker, E.). Oxford: Clarendon Press.

Bancroft, L. (2002). *Why does he do that? Inside the minds of angry and controlling men*. New York: Berkley Books.

Bartky, S. L. (1990). *Femininity and domination: Studies in the phenomenology of oppression*. New York: Routledge.

Blum, L. A. (1982). Kant's and Hegel's moral rationalism: A feminist perspective. *Canadian Journal of Philosophy, XII*(2), 287–302.

Booth, A., Granger, D., Mazur, A., & Kivlighan, K. (2006). Testosterone and social behavior. *Social Forces, 85*(1), 167–175.

Catalano, S (2012) U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Special Report 'Intimate Partner Violence, 1993–2010', NCJ23923. http://bjs.ojp.usdoj.gov/content/pub/pdf/ipv9310.pdf

Cash, M (2002) Distancing Kantian ethics and politics from Kant's Views on Women *Minerva – An Internet Journal of Philosophy* 6. http://www.minerva.mic.ul.ie//vol6/kantian.html

DesAutels, P., & Walker, M. U. (Eds.). (2004). *Moral psychology: Feminist ethics and social theory*. New York: Rowman & Littlefield.

Dillon, R. S. (2012). Critical character theory: Toward a feminist perspective on 'Vice' (and 'Virtue'). In S. S. Crasnow & A. M. Superson (Eds.), *Out from the shadows: Analytical feminist contributions to traditional philosophy*. New York: Oxford University Press.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Gilligan, C. (1987). Moral orientation and moral development. In E. F. Kittay & D. T. Meyes (Eds.), *Women and moral theory*. New York: Rowman & Littlefield.

Horowitz, M. C. (1976). Aristotle and Woman. *Journal of the History of Biology, 9*(2), 183–213.

Johnson, A. G. (2005). The gender knot: Unraveling our patriarchal legacy (Rev and updated ed.). Philadelphia: Temple University Press.

Jordan-Young, R. M. (2010). *Brainstorm: The flaws in the science of sex differences*. Boston: Harvard University Press.

Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral stages: A current formulation and a response to critics*. Basel: Karger.

Mill, J. S. (1869). The subjection of women. London: Longmans, Green, Reader, and Dyer. http://archive.org/details/subjectionofwome00millrich

Stark, E. (2009). *Coercive control: How men entrap women in personal life*. New York: Oxford University Press.

Suhler, C., & Churchland, P. (2011). The neurological basis of morality. In J. Illes & B. J. Sahakian (Eds.), *The Oxford handbook of neuroethics*. New York: Oxford University Press.

U.S. Bureau of Justice Statistics. http://bjs.ojp.usdoj.gov. Retrieved 22 Jan 2013.

U.S. Bureau of Justice Statistics Press Release. (2012). http://bjs.ojp.usdoj.gov/content/pub/press/ipv9310pr.cfm. Retrieved 20 Jan 2013.

A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience

90

Anne J. Jacobson and William Langley

Contents

Abstract

In this chapter, we will consider questions about implicit bias. A case of implicit bias typically involves an agent's action being in conflict with some sincerely professed values and beliefs. Thus, Jones maintains non-Christians may be as honest as Christians, but in practice, Christians are accorded more credibility. Later we will add more features in order to characterize the phenomenon more precisely, but we now have enough to start with.

Implicit bias raises questions in several areas. In **moral theory**, one of the most crucial issues concerns responsibility and guilt. In **psychological studies**, the question of how to substantially moderate or eliminate implicit bias is very important. Critical race theory can employ notions of social groups to bring these two areas together. On the one hand, critical race theory can direct our attention

A.J. Jacobson (✉) • W. Langley
University of Houston Center for Neuro-Engineering and Cognitive Science, Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA
e-mail: Anne.Jacobson@mail.uh.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_157,

away from the facts of individual agency, as opposed to social engagement. From this perspective, culpable wrong-doing comes from complicity in continuing inequalities. A solution is to break down the isolations obtaining among members of different groups.

Recent work from Read Montague's labs can be seen as giving us a neurological model of how implicit bias operates. Further, they have found that expertise protects one against it. We will explore the similarity between the critical race theory's breaking down isolation and the acquisition of expertise some neuroscientists have stressed.

Unfortunately, these similarities do not appear to give us a powerful way to eliminate implicit bias. I conclude by pointing out how these considerations may nonetheless provide a more effective use of some other techniques.

Introduction

This chapter will look at the topic of implicit bias from three perspectives: critical race theory, cognitive neuroscience, and moral theory. In contrast to the standard analytic approach, critical race theory and cognitive neuroscience can be seen to share a view about how to approach implicit bias on matters such as race, gender, disability, sexual orientation, and so on. The approach has some important differences from what we can see as standard philosophy, and those areas of psychology that are congenial to it. For example, it is not wholly individualistic; that is, it does not focus on the individual as the only target for change. Relatedly, it is usually employed with groups of people. Intuitively speaking, using groups can bring in several benefits, among them an increased chance of stability as members reinforce each other's practices. (Examining the effectiveness of forming particular kinds of intentions we will regard as an individualistic approach (Sheeran and Orbell 1999); having people share stores with people of other races is not.) The news is not, however, all good. There remain very serious questions about whether we have discovered a strategy for the sort of change in a society that a significant decrease or elimination of bigotry involves. As we will see at the end of the chapter, the approach can hardly be a complete strategy if our goal is the creation of a more just society.

What Is Implicit Bias?

A paradigm case of implicit attitudes often cited concerns schemas, which we can think of as ways of organizing, and anticipating experience. Our schema for dogs, for example, may include barking and being loyal. We can think of schemas as closely related to stereotypes. Schemas, we can say, are ways of organizing some experiences and stereotypes are the result of using schemas.

For many readers, schemas and stereotypes label familiar phenomenon. A high proportion of philosophers have read Hume on causation and constant conjunctions. In the Anglophone tradition, it may be that almost all of us have. One of Hume's

most basic ideas is that observing type-A things or properties conjoined with type-B leads human beings to feel there is a connection between two such items; in addition, we will expect a B when we observe an A. To take an example related to our discussion, if in one's experience, everyone in a leadership position is male, then one comes to feel there is a connection between the two and to expect a male when one is about to meet a leader (Valian 1999).

I have used the word "feel" in explaining Hume's thought, because it is important that our sense of a connection and our expectations are not the product of reason. They are much more instinctual, for nature, who made us, knows that reason is not a reliable source of these important features of our thought. And making the connections is very important and, for Hume, forms the core of our knowledge.

A great deal of recent cognitive psychology and cognitive neuroscience bears this view out (Montague et al. 1994; McClure et al. 2004). We will in fact make use of one such case, developed by Ann Harvey (Harvey et al. 2010; Kirk et al. 2011). Her work addresses the questions we will shortly raise.

Our treatment of these two topics is not going to come to any easy answer. We will instead accomplish two things. First, we will bring together areas of discourse that still largely exist apart. Secondly, we know that shaming and blaming the perpetrators of bias can very easily backfire; very standard advice is to avoid blaming. At the same time, the prospect of shame and blame can have very dramatic effects on behavior. As we start to turn to non-individualistic pictures, we may well find ways in which shaming and blaming can be employed without their familiar toxic effects.

Bias in Action

One consequence of instinct's role is that it is quite hard not to form the expectations. They are cognitive correlates to bodily instincts, such as blinking when something is in an eye. But, while they are instinctual, they may have profound social impact. Suppose all the professors one has observed are male; the chances that one will select a female to fill a vacancy in the department are slim. One will feel there is a connection between being a professor and being male.

We will focus on biased actions and attitudes that are actually or potentially harmful to the targets of the bias. The biases we will look at are unjust. It is possible to be biased against members of a class for quite good reasons, and actions coming from such a bias may not be unjust; thus, one might be biased against a group's judgments on abortion. Further, we will take attitudes and actions to constitute the bigotry we are interested in when they are accompanied by power differentials. Thus, a white man refusing to hire people of color may be being racist, because in our society, whites have more social power than people of color. In contrast, an average 10-year-old child who does not want to play jump-rope with any adult is not demonstrating ageism, as we will be using standard terms for bigotry.

These scenarios based on stereotypes may play out even when I am unaware of the connection I have made. Thus, despite my conscious view that gender and

intelligence are not connected, I might still, unbeknown to myself, let such associations influence my action.

There is a general assumption in the literature on implicit bias that discoveries in one area of bias (e.g., race or gender orientation) should transfer to the other. Some of the work I will discuss has to do with race, and another large part has to do with one sort of implicit bias money can create. This might seem quite unsatisfactory if the field was more developed, but we are still at a very early stage, when similarities in quite different areas are very much worth investigating.

Our discussion is somewhat focused on racial bias, but it is motivated by a deep concern also for gender bias, and particularly how it plays out in the somewhat unregulated context of academia. In 2005, a report sponsored by the American Historical Association (http://www.historians.org/perspectives/issues/2005/0501/0501new2.cfm) wrote:

> There is more than enough resignation, bitterness, disillusionment, and discouragement to warrant a more serious and extensive consideration of gender in the profession than we were able to carry out in this survey. . . . The profession as a whole should be concerned that so many successful women feel they have suffered from gender discrimination. Female talent is being squandered in fights over large and small issues that could be ameliorated by the attentiveness of administrators, department chairs, and colleagues, and the establishment of more transparent institutional procedures.

It may well be a mistake to take this harassment of women in academia as like that faced by African-Americans, but it can certainly be severe. What can we do about the many people who, because of bias in their society, have their talent squandered?

Responsibility

Bigotry appears to be an attitude of individuals. If we want to consider how to combat it, we need to look the conditions under which individuals are held responsible, when they are. It seems obvious that such a perspective would tell us something about causes, even if less than the whole story. We need to advance a warning, however, which we will return to at the end of this chapter. One remarkable fact about actions prompted by implicit biases is that they are often not the objects of disapproval or blame. Many of them are the result of decisions that do not wear the bigotry on their sleeves. But there is another common reason: Blaming and shaming generally do not work to diminish prejudice and prejudiced actions. People become very defensive and refuse to consider the critic’s point of view. Blaming is not, then, generally cited as one of the weapons we have when fighting bigotry.

Let us now turn to actions. Simplifying somewhat, we can say that the standard philosophical theorists hold a linear causal view of action of a fairly Davidsonian kind; this model may appear even in extended, embodied approaches (Cash 2010). That is, actions are seen as the endpoint of a causal progression from motive to act. One wants a chocolate; one picks a piece up and puts it in one’s mouth. But with implicit prejudice, a serious problem arises immediately. Implicit prejudice

arguably involves ignorance of what one is doing; supposing the chocolate is actually meant to be a present for one's host, ignorance may mean one does not know one is eating some of the host's present. A condition of responsibility is knowledge of what one is doing, and so ignorance may reduce or eliminate responsibility, on the standard philosophical view (though not for all theorists (Holroyd 2012)).

If we consider the question of responsibility abstractly, then it may not seem odd to count bigots as absolved from blame if they are genuinely ignorant of their motives. However, it is difficult to accord innocence to jurors who sentence people to death, because their implicit prejudice leads them to think that people of some ethnicities are not honest. As one would expect, then, critical race theorists are much less enamored of a dismissal of questions of wrong-doing. Their view comes with a different or additional view of action and responsibility. This different view emerges from a number of considerations, one of which will we be considering below.

As we look at models of blameworthy actions, our perspective will change. This should lead us, among other things, to be careful about how we describe what we are doing. A focus on the possession of attitudes by individuals narrows our attention considerably. There are many, many ways in which the more structural factors of a culture may in effect carry bias with them and so induce or re-induce biases. If we are interested in solving a political problem of any magnitude, then we have to include such structures in our targets. In addition, it may be that ameliorating the attitude problem will lessen the chances of effective group political action to solve the structural problems. In light of such concerns, we should look at this chapter so far as directed toward a psychological question, rather than providing a political solution. The psychological question is about how to change attitudes, and not about how more generally to create a more just society.

The Debate About Moral Responsibility

While critical race theorists are not about to excuse whites for judging people of color much more harshly than whites, analytic philosophers have a less easy time saying that the whites are morally responsible. One important obstacle to saying they are morally responsible is, as we have seen, that we tend to think that ignorance provides a very good, firm excuse. If one honestly thinks Mary is a less good student than Joe is, how can one be morally responsible for giving her grades lower than those she merits; that is, ones that are comparable to Joe's?

There are three different, but fairly standard responses from analytic philosophy to the blameworthiness of one's action. All three appear to operate with a similar structure for actions we have looked at. We will start with what we might consider to be a barebones approach that employs a simplified Davidsonian theory, but one that can still be considered the standard theory (Schlosser 2011). This approach understands actions as points in a causal process, ones that start with internal events such as beliefs and desires. Thus, I might desire a piece of chocolate and, checking my purse, find some Hersey kisses that I assume I picked up leaving a restaurant. If it turns out that

they are not mine, that someone else stuck them in my bag by mistake, my mistake about the true origin vitiates any charge of misusing someone else's possession.

A second view, one that makes agency even more demanding, adds in that the action must stem from a self with some unity, and particularly with some unity of one's values. Unconscious elements cannot be part of this unity, and so in a real sense, the action is not one's own action. As a consequence, one is not morally responsible for it (Arpaly and Schroeder 1999; Frankfurt 1971, 2008; Norris 2010).

We can understand the first and second views as potentially excluding responsibility. However, ignorance comes in degrees. Perhaps a friend asked if she could store some candy in my purse; my simply forgetting her request may not be an effective excuse. A third approach may, then, view the matter more as allowing that our judgments of awareness and responsibility may vary from one case to another (Holroyd 2012).

All of these reactions allow the possibility that a sexist grader or a racist juror may be entirely innocent in their unjust grading or sentencing. Nonetheless, we may start to feel our views at odds with this outcome if we look at another case of implicit bias. It turns out that pharmaceutical companies often enough turn up at doctors' offices with free and rather good lunches. This is not done to make sure the medical profession is well nourished. Rather, it has a payoff for the companies, which is that the doctors are considerably more likely to prescribe the medicine the representatives want to promote. At the same time, doctors typically do not think that the lunches bias them in any way. But doctors have a duty to prescribe medicine on more medically relevant grounds than the quality of free lunches they receive. If the lunch has a strong effect on their actions, it is on the cards that there is a failure to fulfill their duty. And typically, we do not think ignorance generally simply excuses such failures.

We can extend the claim. We have a duty, we might say, to form accurate beliefs about other human beings if such beliefs are to affect deeply the distribution of goods in our society and more generally the just treatment members receive. Just as we should query whether we are prescribing the best medicine to patients or giving our students accurate grades, so we have some obligation to query our tendency to think people of our race are more honest and less likely to be criminals. (A somewhat related point occurs in (Shotwell 2011). We might think of ourselves as having a covert checklist, with of the moral quality of an action dependent on how they should be completed.

The checklist fits nicely into the causal model of actions, where, we can see, the guilt accompanying such an action, if there is any, depends on the actions and its antecedents. Nonetheless, there are less obvious and more unattractive effects of such an approach that we need to look at. For one thing, the focus is very limited: Typically, we look just at a person, an action, and an effect, all abstracted from the social context. In addition, the examples foreground the white person as agent, and the minority person is seen as passive recipient. Consequently, they recapitulate one of the central tropes of discrimination.

Critical race theory is more powerful when it takes such a picture as not the only picture of moral quality and, moreover, as positively misguided in some fundamental ways. A key notion for critical race theorists is complicity, as opposed to guilt and

responsibility. Guilt and responsibility are seen in terms of the linear causal model. However, that leaves out the important feature of complicity. Simply being a member of an unjust organization or society strengthens such groups, and makes us complicit in what they are able to accomplish, even when we do not perform directly harmful or prejudicial actions. Looking at the complicity involved in continued membership in an unjust group brings our attention to the group, its problems, and those who suffer or benefit from it. The discrimination people of color suffer is no longer simply a matter of a white person's moral problem (Applebaum 2010).

Along with the notion of complicity go ideas about how to address it. What is common to critical race theory and the cognitive neuroscience we have seen is a set of ideas that I will call a group-based acquisition of expertise. One obvious form of this is the mixed classroom, where white students explore with students of color both the advantages of being white and the disadvantages attaching to a minority position. Such classes are not, however, simply engaged in intellectual enterprises. They are also like encounter or consciousness raising groups, as participants seek a deeper understanding of their biased social setting.

Such training should reflect not just the many dimensions of bigotry and its harms, but also different modes of access to the presence of such bigotry and the potential for its presence. A person with such expertise should be able to do more than correctly label completed actions. They should also usually be able to recognize situations, at least in their normal cultural setting, that are likely to yield discriminatory attitudes or actions, and body language that heralds its display. The expertise we are describing is like those found in a number of other fields. Thus we would expect an expert in twentieth century US art to be able to list various periods in Warhol's work and to recognize them when shown samples.

The comparison between expertise in prejudice and expertise in art is not random. We are about to turn to a neuroscientific investigation of biased judgments about art and how expertise affects the acquisition of such bias.

The Art-Viewing Paradigm, Part One

In the experiment, subjects are asked to judge the quality of paintings they are shown. The paintings are created for the experiment, and not ones the subjects have seen before. Their judgments reveal very marked biases in the subjects, which are due to the context. The pictures are viewed on paper that has a company's logo on it. Though they all get the same amount in the end, the subjects are told that there is a considerable difference between the companies in the amount they are contributing. One company may be sponsoring participants at $300 each, while another only contributes $50 each. Subjects are told that the logos are just randomly assigned, and their size and placement makes no difference.

The subjects overwhelmingly pick the pictures on which the logo of the more generous company appears. Further, there is no awareness on the part of the subjects that their judgments are affected by the differences in contributions. They say things like, "I guess I just have the same judgment as that company had."

Subjects' preferences showed up both in neural activity and in reported preference. When viewing paintings paired with the sponsor's logo, investigators noted increased activity in the ventromedial prefrontal cortex (VMPFC), an area of the brain believed to encode differences in preference. Similar activity was observed in a 2008 experiment in which the effect of price on wine taste preference was tested (Plassmann et al. 2008). Test subjects showed a consistent preference for the "sponsored" paintings in the follow-up behavioral response questionnaire. Researchers also tested for the effect in cases where companies made a "mere offer" of sponsorship. In cases where only one of the two companies offered to sponsor the experiment but neither was selected, subjects still showed a significant preference for paintings paired with the offering company's logo.

The team designing and executing the experiment think the result is due to a fairly deep need on the part of human beings to reciprocate when they receive gifts or benefits. The subjects are in effect giving back to the more generous company. They are not, however, aware of what they are reciprocating.

We have in this case some of the most important features of racial bias, and particularly implicit bias. First, there are the discriminatory judgments. Secondly, we have a hypothesized mechanism. Reciprocity is the problem in the art-viewing case. And while we do not know what the mechanism is for racial bias, it is reasonable to look for some mechanism that has some of the same features reciprocity has, such as being a deep-seated and a widely spread trait. Finally, and most importantly, the subjects are not aware of what is driving their judgments.

The Art-Viewing Paradigm, Part Two

There are all sorts of occasions in which a gift might be given in order to receive a more favorable outcome on some topic. Pharmaceutical companies notoriously are great gift givers. Doctors typically say that a company's free lunches do not affect what they prescribe, but the companies clearly regard it as a strategy worth investing in. Work with the art-viewing paradigm backs the companies' judgment.

If we want to think about the moderation or elimination of bias, the next experiment from the same group gives us a lot of information. The researchers brought together a group of art experts and subjected them to the picture-viewing paradigm, including the information about who was paying what. In the outcome, there was no sign of bias. Expertise appears to protect one against bias (Kirk et al. 2011).

In addition, fMRI studies provided a very revealing difference between the art experts and the naïve subjects. In particular, there were important differences in the brain regions responsible for registering rewards and those for providing for the organism's final reaction to the first. The ventral medial prefrontal cortex (VMPFC) registers a sort of summative view of the rewards available, and in the naïve subjects, it was definitely affected by the pictures and their accompanying logos. The dorsal prefrontal cortex underlies the final judgment on action, and the reciprocity mechanism had created a very definite sense about value for it to pick.

The VMPFC of art experts, on the other hand, was not affected by the difference in reward attached to the accompanying logos. It is not that the DMPFC overcame the VMPFC; rather, the biasing mechanism failed to deliver its standard result. Interestingly enough, a few of the naïve subjects did not give biased judgments about the art, and the VMPFC was similarly less active as the art experts were.

New Problems

The research comparisons looked at so far do not provide us with many *definite* answers, but they raise some interesting questions. For example, can we find some similar bias-generating mechanism in the case of racism, sexism, etc.? Is the protective expertise in the two cases sufficiently similar to ground substantive generalizations from one class to the others. And, especially, are we seeing anything like a way to free us and our society of the scourge of racism?

We will start with an objection that is not as powerful as it may at first sight seem. Then we will move to a recent one that is potentially deadly. Finding out how to combat it will be illuminating.

We can think of the investigation so far as neuro-psychological. It takes the mental operations of individuals as the proper topic; its central question is about changing or modifying them. Our inquiry so far can be contrasted with an entirely different approach, that of collective action. The Civil Rights movement in the United States was a collective attempt to change structures in the society, in contrast to simply creating changes in attitudes toward Black persons. What we turn to now could be said to be more a sociological inquiry into whether what we have been looking at can result in changes in societies.

At least on the face of it, the prospects for change may not look very good. One problem is easily discerned. Given that racism affects nearly all of us, the acquisition of expertise is too limited to match the spread of bigotry in the society. In addition, some of the force of art expertise may well be due to the fact that art experts are to a considerable extent self-selected. Experts entering Montague's labs already have some idea of what discerning artistic judgment is; we have little reason to think that something similar holds in the naive cases.

One response to this problem is to see the training as the development of a select class, as opposed to giving us a potentially general group that can or should include us all. The key here is that expertise has a place in our society; given what people are willing to spend for expertise, it seems to have some initial credibility. Experts can move others to try to share some of their perspective. What we learn, it seems, is that while the acquisition of expertise about white privilege and black deprivation is not a universal solution, it is a promising addition to our discourse.

We will next look at some difficulties raised very recently in John Dixon et al (Dixon 2012a, b). There are two features of their discussion that are different from ours, and indeed from most discussions in philosophy. First of all, they take prejudice to be a matter of, in some general sense, not liking the targets of bigotry. Nonetheless, they seem aware that bigotry need not involve negative attitudes; for example, sexism

need not involve negative feelings about women. Hence, this difference seems not to impact our discussion very much. A second different is more significant. Their attention is not restricted to implicit attitudes. They include many cases, such as those of slave owners, which may not involve unconscious dislike at all; the dislike may be fully conscious. This difference is significant, since presumably, if we cannot put to rest conscious bigotry by some technique or procedure, we can expect unconscious bigotry to be even harder to modify or eliminate.

In discussing Dixon et al's arguments, we will start with evidence against what we have treated as a central claim of critical race theory, namely, that getting groups of people in a perspective sharing cluster will mitigate bigotry. Dixon et al take it as a matter of empirical fact that this does not necessarily work. They focus particularly on Dividio's research, which originally supported perspective sharing, but later questioned it (Hehman et al. 2012). They also look at actual social changes some countries have undergone. Forming such groups, even when they include perspective sharing, may not produce a significantly more egalitarian group. As with male and female relationships, the more powerful may want to preserve their position of power. The less powerful may remain resentful, while at the same time losing their sense that extensive change is needed. This result needs stressing since the relevant research has concentrated more on the attitudes of the oppressors than of the oppressed. Looked at simply from the dominant group's perspective, the success may seem more significant than it really is. The oppressed may be much less impressed.

Dixon's work makes a strong case that social institutions are not eliminated or even changed very much by making white people aware of complicity and white privilege. Historically, change springs from the discontent of those oppressed. Mitigating bias may in fact moderate the discontent enough to weaken a movement toward change; the good may genuinely be an enemy of the better.

For Dixon et al, there is a very important contrast then between a psychological change in individuals and a social change brought about through group action. Their strong case for this conclusion is very worrying, since it means that dealing with bias will be much more intractable than it may have seemed. There is, however, one more approach, one that is somewhat in between the psychological and the sociological.

Discussions of bias tend not to spend a great deal of time on another potentially moderating technique. While some bigotry is very visible in both its operations and its effects, a great deal operates outside of the public gaze. This is obviously true of much in the assessments in academia, yet these relatively small assessments may provide extremely important support of the very visible prejudice. Thus, if Blacks typically do not do well in academia, they may well end up in inferior jobs.

A second area where bigotry can operate quite quietly is health care. African-Americans in general receive inferior care across a lot of dimensions, from pain medication to investigative techniques and onto life-saving surgery. But the underlying decisions are relegated to a fairly private sphere occupied by doctor and patient. Hence, the discrepancy in healthcare between whites and blacks did not become the focus of a systematic study before the twenty-first century.

It is not clear why this is so, but it means that biased behavior does not receive the criticism it should. In fact, there is a good reason for holding back on the criticism: It tends not to do much good. People become highly defensive and inquiry tends to shut down. But if we concentrate on how structures in the society encode bigotry and criticize and change them, we may have a more promising route. We refrain from criticizing individuals and instead work on public critiques of more institutional facts.

To some extent, such mechanisms have been introduced in some areas. For example, it may be unlawful to withhold medical help on the basis of race. Many jobs forbid accepting gifts. Nonetheless, looking at a case with much more discussion than change, we can see that Oxford and Cambridge continue to have a very high proportion of students from the upper socio-economic groups. Hence, blaming and shaming something like possible institutional behavior may offer some hope, but it may need to occur in a context in which people are willing to learn how to do things differently. The possibility of future blaming and shaming may motivate learning, just as the censure on presents stops people from accepting them, at least when they think they could be found out. Nonetheless, while a group may be powerful, members of the group may not feel powerful enough to be happy to give up places in jobs and schools to members of the subordinate race. It is very hard to get such change genuinely welcomed.

Until we have more successful general social change, it seems we need to go back to where we started. We need to improve attitudes and welcome the relatively small changes they give us. Here we may find that knowledge really does make subjects much less naïve, as Harvey's work indicates. At least in some areas, we can make some progress, even if it is not as much as we would like. The coincidence between critical race theory and cognitive neuroscience may also be very useful.

Cross-References

▶ Brain Research on Morality and Cognition
▶ Consciousness and Agency
▶ Ethics of Implicit Persuasion in Pharmaceutical Advertising
▶ Feminist Ethics and Neuroethics
▶ Feminist Neuroethics: Introduction
▶ Feminist Philosophy of Science and Neuroethics
▶ Free Will and Experimental Philosophy: An Intervention
▶ Justice: A Neuroanthropological Account
▶ Moral Cognition: Introduction
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play
▶ Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences
▶ No Excuses: Performance Mistakes in Morality
▶ Psychology and the Aims of Normative Ethics
▶ Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform
▶ The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering

References

Applebaum, B. (2010). *Being white, being good: White complicity, white moral responsibility, and social justice pedagogy*. Lanham: Lexington Books.

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 93*(2), 161–188.

Cash, M. (2010). Extended cognition, personal responsibility, and relational autonomy. *Phenomenology and the Cognitive Sciences, 9*(4), 645–671.

Dixon, J., Levine, M., Reicher, S., & Durrheim, K. (2012a). Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution? *The Behavioral and Brain Sciences, 35*(6), 411–466. doi:10.1017/s0140525x11002214. [Article].

Dixon, J., Levine, M., Reicher, S., & Durrheim, K. (2012b). Beyond prejudice: Relational inequality, collective action, and social change revisited. *The Behavioral and Brain Sciences, 35*(6), 451–459. doi:10.1017/s0140525x12001550. [Article].

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20.

Frankfurt, H. (2008). Inadvertence and responsibility. *The Amherst Lecture in Philosophy, 3*, 1–15. http://www.amherstlecture.org/frankfurt2008/

Harvey, A. H., Kirk, U., Denfield, G. H., & Montague, P. R. (2010). Monetary favors and their influence on neural responses and revealed preference. *Journal of Neuroscience, 30*(28), 9597–9602.

Hehman, E., Gaertner, S. L., Dovidio, J. F., Mania, E. W., Guerra, R., Wilson, D. C., & Friel, B. M. (2012). Group Status Drives Majority and Minority Integration Preferences. *Psychological Science (Sage Publications Inc.), 23*(1), 46–52. doi:10.1177/0956797611423547. [Article].

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*(3), 274–306.

Kirk, U., Harvey, A., & Montague, P. (2011). Domain expertise insulates against judgment bias by monetary favors through a modulation of ventromedial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 108*(25), 10332–10336.

McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., & Montague, P. R. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron, 44*(2), 379.

Montague, P., Dayan, P., Person, C., & Sejnowski, T. J. (1994). Bee foraging in uncertain environments using predictive hebbian learning. *Nature, 377*, 725–728.

Norris, C. (2010). Frankfurt on second-order desires and the concept of a person. *Prolegomena: Casopis za filozofiju, 9*(2), 199–242.

Plassmann, H., et al. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *PNAS, 105*(3), 1050–1054.

Schlosser, M. E. (2011). Agency, ownership and the standard theory. In J. H. Aguilar, A. A. Buckareff, & K. Frankish (Eds.), *New waves in philosophy of action*. Houndmills/Basingstoke/Hampshire/New York: Palgrave Macmillan.

Sheeran, P., & Orbell, S. (1999). Implementation intentions and repeated behaviour: Augmenting the predictive validity of the theory of planned behaviour. *European Journal of Social Psychology, 29*(2–3), 349–369. doi:10.1002/(sici)1099-0992(199903/05)29:2/3<349::aid-ejsp931>3.0.co;2-y.

Shotwell, A. (2011). *Knowing otherwise: Race, gender, and implicit understanding*. University Park: Pennsylvania State University Press.

Valian, V. (1999). *Why so slow? The advancement of women* (1st MIT Press paperback ed.). Cambridge, MA: MIT Press.

Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform

91

Cordelia Fine and Fiona Fidler

Contents

Abstract

Towards the end of the last century, statistical reporting in medical research underwent substantial reform, with null hypothesis significance testing replaced with an estimation approach. Interestingly, this reform may have been largely motivated by the social costs of error within medical research, rather than simply scientific error per se. This chapter briefly reviews the benefits of the estimation statistical approach as a means to producing reliable information about nature and then describes how the current statistical method of null hypothesis significance testing specifically contributes to scientific error in sex/gender neuroscience. The potential social harm that can arise from such errors in this area of

C. Fine (✉)
Melbourne School of Psychological Sciences & Melbourne, Business School & Centre for Ethical Leadership, University of Melbourne, Carlton, VIC, Australia
e-mail: c.fine@mbs.edu

F. Fidler
Australian Centre of Excellence for Risk Analysis (ACERA), Environmental Science, School of Botany, University of Melbourne, Carlton, VIC, Australia
e-mail: fidlerfm@unimelb.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_156,

research is then highlighted. It is suggested that sex/gender neuroscience may therefore provide a valuable model to motivate, on ethical grounds, statistical reform within the psychological sciences.

Introduction

Starting in the early 1980s, statistical reporting in medical research underwent substantial reform. Previously, analysis in medical research was dominated by null hypothesis significance testing (NHST), and the interpretation of findings was made primarily in terms of "statistical significance." Following many years of critique and debate, the medical literature then shifted to an estimation approach (effect sizes and measures of their associated uncertainty, e.g., standard errors and/or confidence intervals and, eventually, meta-analysis). The benefits of the latter statistical approach as a way of producing reliable information about nature – the fundamental goal of science – had been comprehensively and compellingly argued for decades prior to the reform. Interestingly, it was the social costs potentially arising from erroneous scientific conclusions in medicine – rather than simply scientific error per se – that most strongly motivated reform within the medical community (see Fidler 2011). Clearly, both false-positive and false-negative errors in medical research can give rise to significant harm, as when a treatment used by physicians is falsely believed to be effective (or falsely believed to be more effective than another treatment), or when a treatment that *is* effective in treating disease is not used because it is falsely believed to be ineffective. As Fidler (2011) has pointed out, scientific error in psychological research tends not to have the same potential scope for social harm as does medicine, in which error can literally be implicated in "life or death" outcomes. This in part may explain why advocates for a similar statistical reform in psychological science have not yet met with any significant success. However, Fidler notes that scientific error in psychological science is not cost-free, socially, and she discusses two examples in which erroneous conclusions – which would have been averted by the use of effect sizes, confidence intervals, and meta-analytic thinking – contributed to large social costs within the domains of employment aptitude testing and learned helplessness research.

One major category of psychological investigation with clear potential for social harm arising from erroneous scientific conclusions is research that makes comparisons between social groups. Erroneous scientific conclusions about the extent and origins of group differences may influence individual- and social-level attitudes, beliefs, behavior, social perception, self-perception, and social policy. This chapter, in keeping with its inclusion with the feminist section of a neuroethics handbook, focuses specifically on neuroscientific comparisons of males and females. However, similar arguments could potentially be made for other social group comparisons common in psychological work, including those based on age, socioeconomic status, and mental health status. In this chapter, it is argued that the current statistical method of NHST contributes to nontrivial social

harm arising from scientific error. This domain of research may therefore provide a valuable model to motivate, on ethical grounds, statistical reform within the psychological sciences.

The chapter begins, however, with a brief overview and contrast of the statistical approaches of NHST versus effect estimation.

Statistical Significance Versus Effect Estimation

In this section we briefly review a few of the well-known criticisms of NHST (for extensive reviews, see Cumming 2012; Kline 2004). We focus specifically on arguments that NHST increases the scope for false-positive and false-negative error and facilitates a conflation of statistical significance with practical or theoretical significance. This requires first clarifying the true meaning of a p value and/or the concept of statistical significance. A p value of less than 0.05 is commonly misinterpreted to mean that there is a less than 5 % probability that the results are "due to chance." In fact, the information provided by $p < 0.05$ is much more specific and less useful: A p value is the probability of observing a particular experimental result, or one more extreme, given that it doesn't actually exist in the world. To interpret p as the total probability of a chance result ignores the fact that p pertains only to the probability of a type I, or false-positive, error, and that it does not give information about the probability of a type II, or false-negative error. While the type I error rate is typically held constant at 5 %, the typical type II error rate may be closer to 50 % (we explain this further below in our discussion of statistical power). It is therefore a nontrivial mistake to assume that p captures the entire error rate.

Probably due in part to this common misconception within the research community as to the meaning of significant results, there is considerable overconfidence regarding the replicability of a significant result (Haller and Krauss 2002; Lai et al. 2012; Oakes 1986). The "replication fallacy" is the false belief that 1-p provides the probability of replicating a statistically significant effect. In fact, p values offer only very vague information about replication. For example, if you obtained $p = 0.02$ in an experiment and then replicated the experiment with new samples, making conservative assumptions, there is an 80 % chance that your next p will fall between 0.0003 and 0.30 (Lai et al. 2012). This extremely wide interval includes everything from "highly significant" (0.0003) to "don't look twice" (0.3). Most researchers severely underestimate this variability, expecting a range roughly half the size of the interval above (Lai et al. 2012). By contrast, reporting effect sizes and their associated uncertainty (e.g., confidence intervals) makes uncertainty explicit by providing a set of plausible values for the population effect. Such an approach does not preclude decisions: Confidence intervals can be used to reject or fail to reject the null hypothesis, when appropriate, by noting whether or not the null is captured. However, the main benefit of presenting results as effect sizes and confidence intervals, rather than test statistics and p values, is that it facilitates "meta-analytic thinking" (Cumming 2012). That is, it encourages us to compare the size of the

effects across studies, and note how the precision of estimates improves with each new data set, as opposed to making dichotomous decisions (significant/nonsignificant, accept/reject) on the basis of single experiments.

The misconception that a p value reveals the probability that a result is due to chance may also contribute to the overlooking of the possibility for false-negative errors. Statistical power is a measure of the probability that a given experiment will detect an effect of a certain magnitude as "statistically significant" if the effect really exists in the world. Statistical power depends on the size of effect (e.g., some difference of interest, a correlation or coefficient), the variability in the sample (e.g., individual differences), design features of the experiment to control for extraneous variation, sample size, and the alpha criterion set for statistical significance. In psychology and neuroscience, small sample sizes, modest effect sizes, and considerable individual differences (high amounts of variability) combine to create low statistical power in many experiments. While the probabilities of false-positive errors (i.e., p values) appear in over 95 % of articles in psychology journals, probabilities relating to false-negative errors (i.e., statements of statistical power) are reported in fewer than 10 % of articles (Cumming et al. 2007). Over the years, several diligent researchers have gone back through the literature and calculated the statistical power of published experiments. From these independent calculations, we know that the average statistical power of experiments to identify the medium-sized effects typical in many fields of psychology is around 50 % (e.g., Rossi 1990; Sedlmeier and Gigerenzer 1989). This means that conducting an average psychology experiment is roughly equivalent to flipping a coin, in terms of whether you get a statistically significant result or not (Hunter 1997). Button and colleagues recently estimated the statistical power of neuroscience studies to be even lower, at between about 8 % and 31 % (Button et al. 2013). Many statistically nonsignificant results are therefore not good evidence of "no effect" or "no difference." By contrast, effect estimation offers immediate information about precision (an equivalent concept to statistical power). A wide interval indicates a lack of precision; a narrower interval, relatively better precision. Confidence intervals contain a set of plausible values for the population effect, so wider intervals rule out fewer values as plausible. In other words, they give a less focused estimate of the effect. This means that studies with poor precision but nontrivial effect sizes are less likely to be misinterpreted as "no effect" (Fidler and Loftus 2009). It is hoped that this style of reporting would also mean that studies reporting trivial differences between groups are less likely to be misinterpreted as important and meaningful simply because they managed to scrape over the hurdle of statistical significance.

Furthermore, an average statistical power of 50 %, combined with many journals' biases towards only publishing statistically significant results, produces a skewed literature. The notorious "file drawer" is the typical fate for studies that failed to reach the significance threshold, even though there may be more of them. Publication bias pushes the number of false positives in the literature far beyond the 5 % rate one would expect from tests that report $p < 0.05$ – an outcome perhaps facilitated by the aforementioned overconfidence on the part of researchers of the replicability of a positive finding. If instead statistical reporting focused on effect sizes and intervals,

journal editors and reviewers – and researchers themselves – might pay greater attention to the precision of their estimates (as measured by interval width) and the size of the effects, rather than to the single probability of a type I error. If there is an equivalent publication bias that arises with estimation, then it is likely to be one that favors larger effect sizes and higher precision (statistical power). It's difficult to see how any such bias could create problems on the scale as those we see from the current bias towards statistical significance. In fact, it would seem to be exactly the sort of "bias" that many sciences desperately need.

There is a converse issue to the point that a nonsignificant p value is consistent with a sizeable true effect. This is that a significant p value does not guarantee a big and/or interesting effect or difference in the results. Because p values are a function of both sample size and effect size, neither can be read directly from a p value. Small, trivial differences can achieve statistical significance in a study with a large sample size and/or tight controls, for just the same reasons that sizeable important differences can fail to achieve statistical significance when the opposite conditions hold. The distinction between statistical significance and practical or theoretical significance is an important one. Practical and theoretical significance can be taken at face value – is the difference or effect big and important enough to make a practical difference in the world, and/or does it usefully advance scientific theory? Such questions belong in the discussion section of any experimental paper. But to enable a discussion of such questions, measures other than the statistical significance need to be reported. Effect sizes and their associated uncertainty intervals, either standard error or confidence intervals, are the commonly recommended alternatives (e.g., the last three editions APA Publication Manual have recommended reporting effect sizes, and the latter two have additionally recommended reporting confidence intervals). Effect sizes focus directly on the size of the difference, answering the "how big?" question, and lead naturally to questions of practical significance, "is a difference that big (or small) important?" and theoretical significance "what impact does a difference of that size have on our theory?"

Others have also pointed out how easily a conflation between statistical significance and practical or theoretical significance can arise. For example, Gigerenzer (1998, p. 5; see also Meehl 1978) identified null hypothesis significance testing as the tool primarily responsible for diminishing scientific theory in psychology, as it provides researchers with no incentive to formulate a quantitative hypothesis. Instead, the researchers' own unquantified hypothesis (the "alternative" hypothesis) is tested against a "nil" null hypothesis of no difference between the means or zero correlation. At best, the alternative hypothesis may be ordinal (that a difference will be *greater or less than* zero). In contrast, a truly quantitative hypothesis would predict, for example, whether a difference would be greater or less by a factor of three or a factor of 20. Theory would guide the specification of those quantities, and the two competing hypotheses would be pitted against each other. It is extremely rare to see such quantitative hypotheses in psychology (Cumming et al. 2007). An institutionalized methodology that does not encourage researchers to think quantitatively about their hypotheses is unhelpful in terms of progression of scientific theory.

The problem above is not limited to psychology and neuroscience, but it is limited to disciplines that over-rely on statistical significance. In sciences that do not, research questions are posed quite differently. For example, a chemist would not ask: “Is the boiling point of this new substance significantly different from zero?” (Cumming and Fidler 2009). Rather, they would start by empirically estimating the boiling point and quantifying the uncertainty of their estimate. The business of subsequent experiments would be to refine the estimates, with each new data set improving the precision. Theory would direct testing quantitative estimates of different boiling points under different conditions.

The issues discussed above concerning false-positive errors, false-negative errors, and the conflation of statistical significance with practical or theoretical significance can be readily applied to sex/gender neuroscience. Each is discussed briefly below.

False-Positive Errors

False-positive results and publication bias are issues in all areas of behavioral science (for recent discussions, see Fanelli 2012; Simmons et al. 2011; Yong 2012). However, it has been argued that this problem is particularly acute for neuroscientific (in particular, functional neuroimaging) investigations of sex/gender difference, and the arguments below draw extensively on previous discussions in Fine (2012b, forthcoming). As noted long ago (Maccoby and Jacklin 1974), false-positive errors are exacerbated in the sex/gender differences field to the extent that male/female comparisons are routinely made. The simplicity and “obviousness” of testing for differences between males and females, together with a publication bias towards positive findings, creates good conditions for false-positive findings of difference to be reported, while true-negative findings are not. There is currently no way of knowing whether researchers who do not report sex comparisons have nonetheless tested for them, and it is not known how common such practices are. However, Kaiser and colleagues (Kaiser et al. 2009, p. 54) have argued that because sex is a primary and ubiquitous social category, classifying participants by sex is a “natural default” and is “seemingly effortless and obvious in brain research.” Wallentin (2009), moreover, has pointed out that functional neuroimaging studies are especially vulnerable to false-positive sex/gender results, due to nuisance variables (like breathing rate and caffeine intake) that affect the imaging signal, and that this is particularly an issue when sample sizes are small. In line with this, Thirion et al. (2007) demonstrated the low reliability of fMRI studies with samples less than 20, due to large intersubject variability.

A good demonstration of these issues is provided by investigations of the long-standing hypothesis that the male brain is more lateralized than the female brain for language processing. A heavily cited early fMRI investigation of this hypothesis gave 19 men and 19 women phonological, orthographic, and semantic processing tasks to perform while being scanned (Shaywitz et al. 1995). Left lateralization of inferior frontal gyrus activity was found in men for phonological processing, but no

such lateralization was observed for women. No sex/gender differences were observed for orthographic or semantic processing. Subsequent data have been mixed and, overall, unsupportive of Shaywitz et al.'s positive finding: Two recent large meta-analyses of functional neuroimaging studies of language lateralization failed to find evidence for sex/gender differences (Sommer et al. 2004, 2008). Further demonstrating the need for skepticism with respect to the reliability of any one finding of a sex/gender difference, particularly when sample sizes are modest, Ihnen and colleagues found that the sex/gender differences in brain activity observed in a group of 13 men and 13 women during language processing failed to generalize to similar language tasks within a second group of 10 men and 10 women. Moreover, identical analyses of the same participants "discovered" brain activation differences between randomly created groups matched on sex, performance, and obvious demographic characteristics (Ihnen et al. 2009).

Thus, although in theory the probability of false-positive errors should remain the same regardless of sample size, a combination of publication bias, data noise, large intersubject variability, and considerable scope for researcher discretion about the construction of dependent variables may mean that, in practice, this is not the case. This is problematic since, due to the expense of conducting such research, small sample sizes are common in functional neuroimaging investigations of sex/gender. A recent analysis of such studies published in 2009 and 2010 found that it was as common for studies to have fewer than 10 participants in each of their groups of interest (in many studies the sexes were further subdivided, for example, by clinical status or experimental condition) as it was to have more than 20 participants in each experimental cell (Fine 2012b). This underlines yet further the importance of meta-analyses in attempting to establish the size of effects, as individual studies, particularly very small ones, should not in themselves be the basis of theory acceptance or rejection.

Reporting effect sizes and uncertainty intervals as the primary analysis does not make the problems of small sizes, noisy data, or large individual differences go away, but it would help make those problems transparent. For example, we are better able to judge when a statistically nonsignificant effect is simply the product of low power when results are presented as effect sizes and intervals rather than p values alone (see Fig. 91.1). We are similarly better equipped to judge when a result is statistically significant, but the best estimate is too close to zero to be of genuine interest. Under conditions of small sizes, high noise, and large individual differences, intervals will usually be extremely wide. This immediately tells us precision (statistical power) is low. Seeing a figure with error bars that span a very wide range of all possible results would, we hope, encourage caution in the interpretation of statistically nonsignificant results. It would draw attention to the fact that although some intervals fail to include a difference of zero (reflecting a statistically significant result), they may stop only a fraction shy of zero and still be extremely wide. The great advantage of confidence intervals then is that they bind together information about type I and type II errors: 5 % is the total error rate (or chance the interval will miss the true value), and the width of the interval can't be left out or ignored in the way that statistical power routinely is.

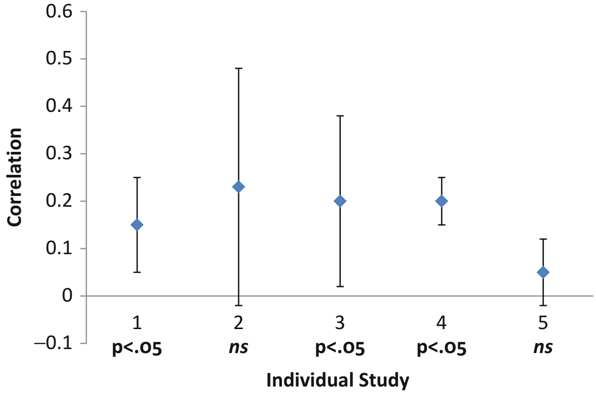

Fig. 91.1 The statistical outcome of five fictional studies comprising a hypothetical research literature. Results of individual studies are reported as correlations, with 95 % confidence intervals. The corresponding NHST result is reported below the horizontal axis. Intervals which cross zero correspond to statistically nonsignificant (*ns*) results at the 5 % level

Imagine the five studies in Fig. 91.1 comprise the existing literature for a question about correlation between biological sex and lateralization of brain activity. Examining only the effect sizes (in this case, correlations) and confidence intervals, the salient message is one of *consistency* of results – the correlations are very nearly the same in each study. There are, however, differences in precision – some intervals are clearly much wider than others. The widest ones (studies 2 and 5) cross zero and are therefore statistically nonsignificant at the 5 % level. If these five studies were reported in terms of statistical significance, then the message would *not* be one of consistency. Rather we see claims of statistical significance (study 1) followed by nonsignificance (2) followed by significance (3 and 4) and nonsignificance again (5). Reviewers of the literature would most likely call for further studies, exacerbating the problem. Debate would be created where in fact none should exist. The current bias in publishing would see study 3 accepted for publication, but study 2 excluded, simply because the latter scrapes over zero and the former narrowly avoids doing so. Study 5 would also be excluded from publication, even though it has relatively high precision compared to others that would be accepted (e.g., 1 and 3). The language of statistical significance allows the bias to keep operating, by creating false dichotomies between, for example, study 2 and 3 which, if they looked at different kinds of language processing, might create a red herring in terms of future research directions. Without a primary focus on *p* values, and the accompanying language of statistical significance, the illusion that any of correlations differ from each other in a meaningful way disappears.

Figure 91.2 shows another set of five fictional studies comprising a hypothetical research literature. In this case, the true population correlation is less than 0.1 and so very unlikely to be theoretically or practically meaningful. Now imagine this is a domain in which publication bias is at play. The nonsignificant studies (studies 2, 3, and 5) would be excluded from publication, even though in every case they have higher statistical power (higher precision, narrower intervals) than the statistically significant studies (studies 1 and 4). The two significant studies leave a slightly inflated view of the correlation, although it remains very small. Even so, if those

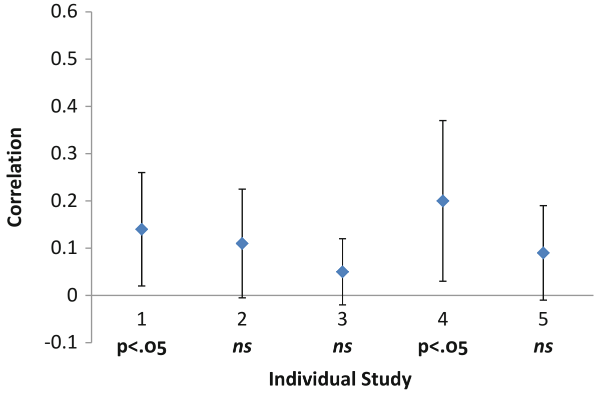

Fig. 91.2 The statistical outcome of five fictional studies comprising a hypothetical research literature, where the true correlation is trivial (i.e., 0.1). The outcome of individual studies is reported with 95 % confidence intervals. The corresponding NHST result is reported below the horizontal axis. Intervals which cross zero correspond to statistically nonsignificant (*ns*) results at the 5 % level

studies are reported primarily in terms of statistical significance, then those small correlations are susceptible to being misinterpreted as meaningful. If instead the actual correlations with their very wide error bars are observed, narrowly avoiding zero, there may be less chance of misinterpretation. There may also be greater awareness of sampling error and suspicion regarding what lies in file drawers and whether replications would also scrape over the line.

In addition, it is not only the production of false-positive errors that is of concern in this area of research – it is also the persistence of false-positive claims in the research literature. An analysis by Fine (2012b) of all 2009 and 2010 citations of the original Shaywitz et al. (1995) finding found that 43 of the 75 (57 %) cited the Shaywitz study without referring to the existence of any contradictory data. Of the remaining 32 citations, 11 (15 % of the total) either failed to cite either of the Sommer and colleagues' meta-analyses or did so in a way that was misleading. A further nine studies cited one or both meta-analyses, but in a way that gave no indication that, through virtue of its status as a meta-analytic study, it lay claim to being a more reliable source of information than Shaywitz et al. Thus, just 12 of the 75 (16 %) studies that cited the Shaywitz finding also cited the work of Sommer and colleagues in a fully informative way (even if the latter's conclusions were disputed). Meta-analysis and the estimation approach go hand in hand. Almost all meta-analysis requires effect sizes of individual studies as their primary input. Furthermore, reporting and interpreting effect sizes and confidence intervals helps researchers think meta-analytically about the contribution of each new study to the existing body of knowledge. They are less likely to see illusory inconsistencies in literatures, such as in the five fictional studies examples (Coulson et al. 2010). For example, one response to Sommer et al. (2004) negative conclusion was the suggestion that males are more lateralized specifically for passive listening to stories (Kitazawa and Kansaku 2005). However, this was made on the basis of two small statistically significant studies using this dependent variable. Such a suggestion might be compelling within the context of a qualitative review that merely assessed significance versus

nonsignificance, but in this case Sommer and colleagues were able to note directly in response that positive findings were much more likely to be observed in small studies (Sommer et al. 2005).

False-Negative Errors

Feminist critiques of neuroscientific investigations of sex/gender difference have been particularly concerned with false-positive errors and the emphasis on difference over similarity (e.g., Kaiser et al. 2009). However, it is also the case that male/female behavior is not completely overlapping in all domains, and the sexes also differ in incidence of mental disorders such as depression and autism. Such differences have, understandably, motivated arguments for the importance of investigating sex differences in the brain (Cahill 2006, 2010; McCarthy et al. 2012). However, the neural basis of such differences will not be readily identifiable in humans. The substantial overlap of male and female distributions on the majority of psychological and behavioral measures has been persuasively argued for quantitatively (Hyde 2005). That male and female brain characteristics are also significantly overlapping, even where statistically significant sex differences have been identified, has also been noted (e.g., Jordan-Young 2010). Moreover, Joel (2011, 2012) has recently argued, on the basis of nonhuman animal neuroscientific research, that although biological sex influences brain development as it occurs in interaction with environmental experiences, it is not in any simple way that generates distinctive "male" and "female" brains. Joel argues instead that brain characteristics take the form of complex, idiosyncratic "gender mosaics" within each individual. This argument highlights the point that, while no doubt the brain is not infinitely malleable, neural circuitry develops through, and is altered by, experience. The importance of this with respect to sex/gender in the brain comes from the fact that gender, as a powerful and pervasive social phenomenon, ensures that a person's biological sex will influence the experiences (such as material, social, and mental) she or he encounters. This will, in turn, leave neurological traces (see, e.g., Fausto-Sterling 2000, 2005; Kaiser 2012 for arguments for the inextricable intertwinement of sex and gender in the body and brain), and such experiences will vary across individuals.

These points raise the important conceptual question of whether it makes sense to try to identity an effect size of biological sex on brain structure or function. But whatever precise research question is pursued, uncovering what are undoubtedly highly complex interactions against a background of noise and considerable individual differences will require more complex experimental designs. As the complexity of design increases, with multiple groups and multiple comparisons, so too must the sample size if adequate statistical power is to be achieved. This would point to the conclusion that the success of future neuroscientific research into sex/gender may depend on the extent to which the field is sensitive to concerns regarding statistical power. Currently, considerable intersubject variability, together with mostly modest behavioral

effect sizes, suggests that typical functional neuroimaging studies may lack the statistical power to detect any true sex/gender effects that may exist – a fact that is obscured by the use of NHST.

Practical and Theoretical Significance

In an analysis of all 2009 and 2010 functional neuroimaging investigations of sex/gender differences, Fine (2012b) found that none of the 39 studies tested predictions derived from well-specified neurocognitive accounts of the sex-/gender-modulated mental processes involved in the behavior of interest. Rather, the studies tested the null hypothesis that no sex differences in brain activity would be observed – either in any part of the brain or in particular regions suggested by prior research. Such a research strategy makes a very limited contribution to theoretical progress, for reasons discussed previously and, in addition, because the functional significance of a sex difference in brain activity is obscure (see Fine 2010; Hoffman 2011). Moreover, the absence of well-developed and empirically supported accounts of the relationship between brain activity and mental processes readily enables interpretations of findings that draw on gender stereotypes to fill in the substantial gaps in neuroscientific knowledge. In the 2009/2010 sample, approximately two-thirds of the studies speculated a functional significance to their finding of sex differences. The influence of gender stereotypes was especially clear where such speculations were consistent with gender stereotypes but inconsistent with the researchers' own relevant behavioral data, which occurred in a number of cases. In other words, there appears not to be a strong requirement for studies to demonstrate either theoretical or practical significance, only statistical significance.

Theoretical and practical significance requires interpretation of results in terms of their effect sizes, not their p values. Such interpretation requires knowledgeable judgment in context, which may be harder than an automated judgment of p relative to an arbitrary criterion, usually.05. But such judgment is accepted as the norm in relation to numerous other aspects of planning and running research; we need to accept it as necessary also for the interpretation of results. Readers may, of course, disagree, but effect size and CI information provide full information for an informed discussion about the results and what they can justifiably imply. That is, one justifies why a 0.6 s reaction time difference or a 5 percentage point difference between males and females in a difference in brain activity between a control and experimental condition is meaningful, large, or important, not why $p = .03$ matters. Any given effect size (e.g., 5 percentage points) may or may not be statistically significant, depending on the other conditions of the experiment, but whether a difference of that size is important in that context will remain constant. Effect sizes underpin theoretical and statistical significance, and confidence intervals give us a plausible range for the effect size. We can then determine whether that range is sufficiently narrow (includes all important and very important effect sizes) or inadequately broad (includes all important effect sizes but also all unimportant ones).

Social Harm from Scientific Error

As argued in Fine (2012a, 2013), it can be expected that claims about sex/gender differences in the brain and their implications for how and why males and females behave differently will have psychological consequences. Hacking (1995, p. 351) has described "looping" or "feedback effects in cognition and culture," whereby the causal understanding of a particular social group changes the very character of the group, leading to further change in causal understanding. Similarly, Choudhury and colleagues, referring specifically to the social impact of neuroscience, have argued that the representation of "brain facts" in the media, policy, and lay perceptions influences society in ways that can affect the very mental phenomena under investigation (Choudhury et al. 2009). The original finding of persuasive power of brain images (McCabe and Castel 2008) has recently been disputed both qualitatively (Farah and Hook 2013) and quantitatively in a recent meta-analysis (Michael et al. 2013). However, "brain facts," regardless of the presence or absence of brain images, may enhance how satisfactory or valuable lay people judge scientific explanations of psychological phenomena to be (Michael et al. 2013; Morton et al. 2006; Weisberg et al. 2008). The enhancing effect of neuroscientific accounts may be facilitated by the way in which neuroscience findings tend to be presented in the popular media (Racine et al. 2005, 2010).

Three lines of research offer insights into the psychological effects of claims about sex differences in the brain and, in particular, implications from such research that gender differences are "essential" (see Fine 2012a). First, preliminary evidence looking at the effects of information about the reasons for gender differences in mathematics on performance suggests that a spontaneous assumption that gender differences are "biologically" caused may have detrimental effects on performance (Dar-Nimrod and Heine 2006; Thoman et al. 2008). Second, both the endorsement of "biological" explanations of gender differences and exposure to such accounts are associated with greater endorsement of gender stereotypes (Brescoll and LaFrance 2004; Martin and Parker 1995) and more stereotypical self-perception (Coleman and Hong 2008). This is merely one example of a general tendency for biological essentialist beliefs to be associated with endorsement of a wide variety of social stereotypes (Bastian and Haslam 2006). Third, there is evidence that a stronger weighting of genetic influence on behavior is associated with greater moral tolerance of the social status quo (Dambrun et al. 2009; Keller 2005). It has been argued that the belief that differences between social groups are biologically essential naturalizes inequality and serves a system-justifying function, both in general (Yzerbyt et al. 1997) and specifically in relation to gender (Bem 1993). Supporting such a motivated system-justifying basis for biological essentialist accounts of gender, Morton and colleagues found links between essentialist beliefs and sexism in men only and only when women were presented as gaining ground on men. Conversely, they also showed that scientific claims about gender influenced endorsement of hierarchy legitimizing beliefs and attitudes (Morton et al. 2009).

In short, erroneous scientific claims are not cost-free. “Brain facts” about sex differences make their way into popular culture where they may influence the very mental phenomena under investigation. In this way, independent of their truth, erroneous claims can affect people’s lives and society in self-fulfilling ways.

Conclusion

There is no “perfect” way to conduct scientific research nor any ideal statistical approaches. In fact, what matters most is sound statistical reasoning, rather than any specific tool. However, for the reasons discussed here, current usage of the single technique of NHST puts scientific reasoning at risk and facilitates the production of research literatures readily contaminated by illusory inconsistencies, false positives, false negatives, and the conflation of statistical and theoretical/practical significance. Other techniques, such as estimation, may promote reasoning simply because they are more transparent and have not been ritualized to the same extent. In this chapter we have provided a specific example of how a replacement of NHST with effect estimation would have a number of beneficial scientific effects within the domain of sex/gender neuroscience. In particular, it would reduce the scope for the production and persistence of false-positive errors, draw attention to the low chance researchers have of identifying real sex/gender effects due to low statistical power, and facilitate more strongly theory-driven research hypotheses. Such changes would not only be scientifically advantageous but also reduce the potential for social harm arising from erroneous scientific claims about sex differences in the brain. That the internal goal of science is not being well served by current neuroscientific investigation of sex differences should create consternation in the scientific community even in the absence of social concerns about harmful effects of erroneous claims about sex differences in the brain. Scientists are certainly responsible for the quality of the evidence they produce, and their research decisions – including statistical approach – impact the probability of error. We suggest here that scientists may also wish to consider the social costs of potential error. The example of medical research suggests that it is these social consequences that, in the end, may be the most persuasive in bringing about beneficial statistical reform that will, in turn, help the progress of science.

Cross-References

▶ A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience
▶ Feminist Ethics and Neuroethics
▶ Feminist Neuroethics: Introduction
▶ Feminist Philosophy of Science and Neuroethics
▶ Neuroethics and Identity

References

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology, 42*, 228–235.

Bem, S. (1993). *The lenses of gender: Transforming the debate on sexual inequality*. New Haven: Yale University Press.

Brescoll, V., & LaFrance, M. (2004). The correlates and consequences of newspaper reports of research on sex differences. *Psychological Science, 15*(8), 515–520.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376.

Cahill, L. (2006). Why sex matters for neuroscience. *Nature Review Neuroscience, 7*(6), 477–484.

Cahill, L. (2010). Sex influences on brain and emotional memory: The burden of proof has shifted. In I. Savic (Ed.), *Sex differences in the human brain, their underpinnings and implications* (Vol. 186, pp. 29–40). Amsterdam: Elsevier.

Choudhury, S., Nagel, S., & Slaby, J. (2009). Critical neuroscience: Linking neuroscience and society through critical practice. *BioSocieties, 4*, 61–77.

Coleman, J., & Hong, Y.-Y. (2008). Beyond nature and nurture: The influence of lay gender theories on self-stereotyping. *Self and Identity, 7*(1), 34–53.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement, 1*.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology, 217*, 15–26.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230–232.

Dambrun, M., Kamiejski, R., Haddadi, N., & Duarte, S. (2009). Why does social dominance orientation decrease with university exposure to the social sciences? The impact of institutional socialization and the mediating role of "geneticism". *European Journal of Social Psychology, 39*, 88–100.

Dar-Nimrod, I., & Heine, S. (2006). Exposure to scientific theories affects women's math performance. *Science, 314*, 435.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904.

Farah, M. J., & Hook, C. J. (2013). The seductive allure of "Seductive Allure". *Perspectives on Psychological Science, 8*(1), 88–90.

Fausto-Sterling, A. (2000). *Sexing the body: Gender politics and the construction of sexuality*. New York: Basic Books.

Fausto-Sterling, A. (2005). The bare bones of sex: Part 1–sex and gender. *Signs: Journal of Women in Culture and Society, 30*(2), 1491–1527.

Fidler, F. (2011). Ethics and statistical reform: Lessons from medicine. In A. T. Panter & S. K. Sterba (Eds.), *Handbook of ethics in quantitative methodology*. New York: Routledge.

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie/Journal of Psychology, 217*, 27–37.

Fine, C. (2010). *Delusions of gender: How our minds, society, and neurosexism create difference*. New York: WW Norton.

Fine, C. (2012a). Explaining, or sustaining, the status quo? The potentially self-fulfilling effects of 'hardwired' accounts of sex differences. *Neuroethics, 5*(3), 285–294.

Fine, C. (2012b). Is there neurosexism in functional neuroimaging investigations of sex differences? *Neuroethics, 6*(2), 369–409.

Fine, C. (2013). Neurosexism in functional neuroimaging: From scanner to pseudo-science to psyche. In M. Ryan & N. Branscombe (Eds.), *The Sage handbook of gender and psychology*. Thousand Oaks, CA: Sage.

Gigerenzer, G. (1998). Surrogates for theory. *Theory & Psychology, 8*, 195–204.

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. Premack (Eds.), *Causal cognition: A multidisciplinary approach* (pp. 351–383). Oxford: Oxford University Press.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*, 1–20.

Hoffman, G. (2011). What, if anything, can neuroscience tell us about gender differences? In R. Bluhm, A. Jacobson, & H. Maibom (Eds.), *Neurofeminism: Issues at the intersection of feminist theory and cognitive science*. Basingstoke: Palgrave Macmillan.

Hunter, J. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 3–7.

Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592.

Ihnen, S. K. Z., Church, J. A., Petersen, S. E., & Schlaggar, B. L. (2009). Lack of generalizability of sex differences in the fMRI BOLD activity associated with language processing in adults. *NeuroImage, 45*(3), 1020–1032.

Joel, D. (2011). Male or female? Brains are intersex. *Frontiers in Integrative Neuroscience, 5*, 57.

Joel, D. (2012). Genetic-gonadal-genitals sex (3G-sex) and the misconception of brain and gender, or, why 3G-males and 3G-females have intersex brain and intersex gender. *Biology of Sex Differences, 3*(1), 27.

Jordan-Young, R. (2010). *Brain storm: The flaws in the science of sex differences*. Cambridge, MA: Harvard University Press.

Kaiser, A. (2012). Re-conceptualizing "sex" and "gender" in the human brain. *Zeitschrift für Psychologie/Journal of Psychology, 220*(2), 130–136.

Kaiser, A., Haller, S., Schmitz, S., & Nitsch, C. (2009). On sex/gender related similarities and differences in fMRI language research. *Brain Research Reviews, 61*(2), 49–59.

Keller, J. (2005). In genes we trust: The biological component of psychological essentialism and its relationship to mechanisms of motivated social cognition. *Journal of Personality and Social Psychology, 88*(4), 686–702.

Kitazawa, S., & Kansaku, K. (2005). Sex difference in language lateralization may be task-dependent. *Brain, 128*(5), E30.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Lai, J., Fidler, F., & Cumming, G. (2012). Subjective p intervals. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*(2), 51–62.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford: Stanford University Press.

Martin, C., & Parker, S. (1995). Folk theories about sex and race differences. *Personality and Social Psychology Bulletin, 21*(1), 45–57.

McCabe, D., & Castel, A. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*, 343–352.

McCarthy, M., Arnold, A., Ball, G., Blaustein, J., & De Vries, G. J. (2012). Sex differences in the brain: The not so inconvenient truth. *Journal of Neuroscience, 32*(7), 2241–2247.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Michael, R., Newman, E., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non)persuasive power of a brain image. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-013-0391-6.

Morton, T., Haslam, S., Postmes, T., & Ryan, M. (2006). We value what values us: The appeal of identity-affirming science. *Political Psychology, 27*(6), 823–838.

Morton, T., Haslam, S., & Hornsey, M. (2009). Theorizing gender in the face of social change: Is there anything essential about essentialism? *Journal of Personality and Social Psychology, 96*(3), 653–664.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester: Wiley.
Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews Neuroscience, 6*(2), 159–164.
Racine, E., Waldman, S., Rosenberg, J., & Illes, J. (2010). Contemporary neuroscience in the media. *Social Science & Medicine, 71*(4), 725–733.
Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.
Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–315.
Shaywitz, B., Shaywitz, S., Pugh, K., Constable, R., Skudlarski, P., Fulbright, R., et al. (1995). Sex differences in the functional organization of the brain for language. *Nature, 373*, 607–609.
Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.
Sommer, I., Aleman, A., Bouma, A., & Kahn, R. (2004). Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. *Brain, 127*, 1845–1852.
Sommer, I., Aleman, A., & Kahn, R. S. (2005). Size *does* count: A reply to Kitazawa and Kansaku. *Brain, 128*, E31.
Sommer, I., Aleman, A., Somers, M., Boks, M. P., & Kahn, R. S. (2008). Sex differences in handedness, asymmetry of the Planum Temporale and functional language lateralization. *Brain Research, 1206*, 76–88.
Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage, 35*(1), 105–120.
Thoman, D., White, P., Yamawaki, N., & Koishi, H. (2008). Variations of gender-math stereotype content affect women's vulnerability to stereotype threat. *Sex Roles, 58*, 702–712.
Wallentin, M. (2009). Putative sex differences in verbal abilities and language cortex: A critical review. *Brain and Language, 108*(3), 175–183.
Weisberg, D., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477.
Yong, E. (2012). Bad copy. *Nature, 485*, 298–300.
Yzerbyt, V., Rocher, S., & Schadron, G. (1997). A subjective essentialist view of group perception. In R. Spears, P. Oakes, N. Ellemers, & S. A. Haslam (Eds.), *The social psychology of stereotyping and group life* (pp. 20–50). Oxford: Blackwell.

Section XIX

Neuroscience, Neuroethics, and the Media

Neuroscience, Neuroethics, and the Media

92

Eric Racine

Contents

Abstract

Contemporary neuroscience has become a poster child for medical sciences and has gained tremendous salience in the public eye. The growth of neuroscience has resulted in multiple developments, generating a flow of new research, allowing the exploration of new research domains beyond the traditional frontiers of medical science, and enticing the younger generations to train in neuroscience. At the same time, the public, in its plural manifestations, has become eager to engage with the advances of neuroscience. In this introduction to this section of the Handbook, I underscore and discuss briefly three aspects of the relationship between the media and neuroethics, which illustrate the importance of the media. I then introduce three original contributions dealing with the media and neuroethics. Hopefully, readers of this book will find these contributions stimulating and that the latter will encourage more needed work in this area.

E. Racine
Neuroethics Research Unit, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

Department of Medicine and Department of Social and Preventive Medicine, Université de Montréal, Montréal, QC, Canada

Departments of Neurology and Neurosurgery, Experimental Medicine & Biomedical Ethics Unit, McGill University, Montréal, QC, Canada
e-mail: eric.racine@ircm.qc.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_82,

Introduction

In many ways, contemporary neuroscience has become a poster child for medical sciences and has gained tremendous salience in the public eye. Starting in the 1960s and 1970s, the international neuroscience community has grown almost steadily. International societies such as the Society for Neuroscience and the International Brain Organization have seen their memberships reach considerable proportions. This growth has resulted in multiple developments, such as generating a flow of new research, allowing the exploration of new research domains beyond the traditional frontiers of medical science, and enticing the younger generations to train in neuroscience. At the same time, the public, in its plural manifestations, has become eager to engage with the advances of neuroscience. A cascade of contributions written for the general public has approached the diagnosis and treatment of neurological and psychiatric illness as well as explored the application of neuroscience knowledge to education, law, and the psychology of everyday life (e.g., creativity, parenting, memory). It is therefore no surprise that the public salience of neuroscience has intimate and multifaceted connections to the field of neuroethics: As neuroscience ventured closer and closer into practical and daily life, it has generated questions about its impact on human values. In the following, I underscore and discuss briefly three aspects of the relationship between the media and neuroethics to introduce this section of the *Handbook*.

Public Notoriety and Visibility of Neuroscience as a Precondition for Neuroethics

First, neuroethics as we now know it would not have emerged as a distinct endeavor were it not for the public notoriety of neuroscience itself. (This does not mean that ethical challenges in neuroscience would not exist without neuroethics or that other responses – other than neuroethics – to these challenges would have not flourished). Scholars and others are keenly aware that advances in neuroscience have percolated to the broader public and public domain, and are therefore also aware of their potential implications. Likewise, ethicists and others have less interest in working on topics which have no bearing on public opinion and behavior (unlike the use of neuropharmaceuticals for enhancement of performance, or non-referred purchase of neuroimaging services), which touch the public directly.

Neuroethics as a Step in the Development of a More Open and Public Bioethics

Second, the field of neuroethics emerged at a time when bioethics itself was opening to empirical research and reconnecting with an impetus found at the beginning of its history, i.e., to consider non-expert perspectives (e.g., see chapter of the Belmont Report dealing with public attitudes toward science and Racine (2010) for a review). Because the impact of neuroscience was potentially far-reaching, there were early

calls in neuroethics supporting the need for public dialogue, public engagement, and general increased public understanding of neuroscience (Safire 2002; Leshner 2005). Arguments were made that the challenge of the interpretation of neuroscience findings, especially those bearing on cognitive processes and personality, brought a conflict between the manifest and the scientific images of man (read: humankind) following Sellars' distinction (Sellars 1963). The public's perspectives and interpretations in the forms of neuroessentialism or neurodeterminism are therefore of importance, since they capture an interpretation of how the manifest and scientific images are negotiated (Illes and Racine 2005). Whether these interpretations are justified or not, or based on solid neuroscience and sound philosophy, the public's deciphering of them would matter as such and would need to be considered.

Neuroethics as a Hub of Research on Media, Science, and Public Engagement

Third, and in response to calls for public engagement, neuroethics has given considerable attention to topics related to public engagement, public understanding of neuroscience, and the media. In a review focused on the early years of neuroethics, this topic was found to be frequent in peer review literature and also in print media reports about neuroethics (Racine 2010). The number of relevant papers is of course too numerous to list or review here (a more comprehensive review has been offered elsewhere (Racine 2011)), but a few important results illustrate domains where progress has been made (see Table 92.1).

Contributions on Neuroscience and the Media

The following contributions in this *Handbook of Neuroethics* by Forlini and colleagues, Krahn, and Savane offer a clear representation of recent work in the area of media and neuroethics and address some of the points discussed above. Forlini and colleagues in a chapter entitled ▶ "Popular Media and Bioethics Scholarship: Sharing Responsibility for Portrayals of Cognitive Enhancement with Prescription Medications" (Chap. 93) approach the roles of the media in debates about cognitive enhancement using prescription medications. They examine how the media may have had upstream implications on bioethics scholarship as well as downstream consequences (e.g., on prevalence rates for the use of prescription medications). Their analysis highlights some problematic aspects of media coverage of cognitive enhancement and how some of these trends have percolated to academic scholarship without sufficient scrutiny. They call for greater awareness of the impact that the media and bioethics scholarship have on such debates and identify questions to address with respect to conventional and social media in the context of cognitive enhancement.

Krahn in his chapter titled ▶ "Traumatic Brain Injury and the Use of Documentary Narrative Media to Redress Social Stigma" (Chap. 95) introduces a critical note about how neuroethics is unfortunately often depicted narrowly by its critiques to exclude

Table 92.1 Illustrative (non-exhaustive) areas of research related to neuroethics in media

Media content dissemination and public portrayal of neuroethical issues
Examination of neuroethical questions and topics such as cognitive enhancement (Coveney et al. 2008; Williams et al. 2008; Forlini and Racine 2009; Partridge et al. 2011)
Examination of media portrayal of neuroscience technologies and neuroscience results such as fMRI and DBS research and related ethical discussion (Racine et al. 2006, 2007a, 2010; O'Connell et al. 2011)
Examination of media content regarding ethical aspects in neurological disorders and mental illness such as coma (Wijdicks and Wijdicks 2006a, b) and the vegetative state (Racine et al. 2008; Striano et al. 2009)
Media impact and understanding of behavior regarding neuroethical issues
Examination of practices and regulatory issues surrounding direct-to-consumer advertising regarding neuroimaging, dietary supplements, and neuropharmacology (Illes et al. 2003; Illes et al. 2004; Racine et al. 2007b)
Examination of the impact of media content on attitudes toward cognitive enhancement (Forlini and Racine 2012)
Examination of the impact of neuroessentialism and neurorealism on neuroscience explanations (McCabe and Castel 2008; Vohs and Schooler 2008; Weisberg et al. 2008)
Nonconventional media and multidirectional communication in neuroethics
Proposal of models for public engagement regarding neuroethical issues (Blakemore 2002; Rose 2003; Racine et al. 2005)
Experiences in multidirectional communication (Illes et al. 2005)
Proposal for training programs in neuroscience communication with relevance to neuroethics (Illes et al. 2010)

considerations related to culture. Inspired by a pragmatic view of neuroethics where public and intercultural neuroethics hold major importance, he offers an insightful analysis of stigma and public misunderstandings in the context of traumatic brain injury where "hidden disabilities" and the "brain" aspect of the injury and related impairments can augment stigma. Krahn then explores a series of novel approaches such as narrative-based strategies and new media to combat sigma.

Savane presents in her chapter, appropriately titled ▶ "Neuroethics Beyond Traditional Media" (Chap. 94), an overview of nonconventional media and public engagement. She focuses on the rich and diverse European neuroscience environment which has witnessed a series of national and European-level initiatives such as *Brains in Dialogue* and *Meeting of the Minds*. These initiatives and the discussion by Savane invite us to consider how public engagement can be part of every aspect of neuroethics: from defining and understanding the issues, to envisioning responses. Although multidirectional communication and public engagement initiatives do not replace traditional media, they respond to a need to broaden public discussion and allow the mutual enrichment of perspectives.

Future Directions

The chapters on media and public engagement in this book meaningfully contribute to the critical investigation and discussion of media in neuroethics and

neuroscience. Still, there are important questions which have been untapped in our field. For example, Table 92.1 presented above is silent on the process of “media content development.” This is an essential area in media studies that has not been fully considered in the context of neuroethics. For example, how is media coverage of neuroethics developed? How do neuroethicists engage with the media and why? How do journalists understand neuroethics and what prompts them to write about these topics? Who gets to talk and obtain coverage and who does not? How are decisions to include or exclude ethical considerations from the reporting of neuroscience research made? These are questions that would call for further attention to understand how the field of neuroethics can contribute to an enlightened public dialogue and a strong representation of matters of ethics in public discussions. Media analysis and discussions about public engagement have been rather well represented in neuroethics scholarship; yet, there is much to be done to push the field further both in scope and depth. In terms of scope, we need replication studies and further comparative work to build on initial studies. In terms of depth, we need to better understand, based on theoretical and empirical contributions, the full cycle of media information development and its relationship to neuroethics. This would involve approaches such as normative models clarifying the responsibilities of neuroscientists in public communication, conceptual clarifications of the impact of public information on neuroethical issues, and tighter empirical work to understand the precise impact of public information on behavior. I hope that readers of this book will find the following contributions stimulating and that the latter will encourage more work in this area.

Acknowledgments The author would like to thank Mr. John Aspler and Mrs. Allison Yan for feedback and assistance on a previous version of this manuscript. Support for the writing of this introduction comes from the Canadian Institutes of Health Research (New Investigator Award) and the Social Sciences and Humanities Research Council of Canada.

Cross-References

▶ Neuroethics Beyond Traditional Media
▶ Popular Media and Bioethics Scholarship: Sharing Responsibility for Portrayals of Cognitive Enhancement with Prescription Medications
▶ Traumatic Brain Injury and the Use of Documentary Narrative Media to Redress Social Stigma

References

Blakemore, C. (2002). *From the “public understanding of science” to scientists’ understanding of the public*. Paper presented at the neuroethics conference: Mapping the field, San Francisco.

Coveney, C. M., Nerlich, B., & Martin, P. (2008). Modafinil in the media: Metaphors, medicalisation and the body. *Social Science and Medicine, 68*(3), 487–495.

Forlini, C., & Racine, E. (2009). Disagreements with implications: Diverging discourses on the ethics of non-medical use of methylphenidate for performance enhancement. *BMC Medical Ethics, 10.*

Forlini, C., & Racine, E. (2012). Added stakeholders, added value(s) to the cognitive enhancement debate: Are academic discourse and professional policies sidestepping values of stakeholders? *AJOB Primary Research, 3*(1), 33–47.

Illes, J., & Racine, E. (2005). Imaging or imagining? A neuroethics challenge informed by genetics. *American Journal of Bioethics, 5*(2), 5–18.

Illes, J., Fan, E., Koenig, B., Raffin, T. A., Kann, D., & Atlas, S. W. (2003). Self-referred whole-body CT imaging: Current implications for health care consumers. *Radiology, 228*(2), 346–351.

Illes, J., Kann, D., Karetsky, K., Letourneau, P., Raffin, T. A., Schraedley-Desmond, P., et al. (2004). Advertising, patient decision making, and self-referral for computed tomographic and magnetic resonance imaging. *Archives of Internal Medicine, 164*(22), 2415–2419.

Illes, J., Blakemore, C., Hansson, M. G., Hensch, T. K., Leshner, A., Maestro, G., et al. (2005). International perspectives on engaging the public in neuroscience. *Nature Reviews Neuroscience, 6*(12), 977–982.

Illes, J., Moser, M. A., McCormick, J. B., Racine, E., Blakeslee, S., Caplan, A., et al. (2010). Neurotalk: Improving the communication of neuroscience research. *Nature Reviews Neuroscience, 11*(1), 61–69.

Leshner, A. I. (2005). It's time to go public with neuroethics [Editorial]. *American Journal of Bioethics, 5*(2), 1–2.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*(1), 343–352.

O'Connell, G., De Wilde, J., Haley, J., Shuler, K., Schafer, B., Sandercock, P., et al. (2011). The brain, the science and the media. The legal, corporate, social and security implications of neuroimaging and the impact of media coverage. *EMBO Reports, 12*(7), 630–636.

Partridge, B., Bell, S., Lucke, J., Yeates, S., & Hall, W. (2011). Smart drugs "as common as coffee": Media hype about neuroenhancement. *PloS One, 6*(11), e28416.

Racine, E. (2010). *Pragmatic neuroethics: Improving treatment and understanding of the mind-brain*. Cambridge, MA: MIT Press.

Racine, E. (2011). Neuroscience and the media: Ethical challenges and opportunities. In J. Illes & B. Sahakian (Eds.), *Oxford handbook of neuroethics* (pp. 783–802). Oxford: Oxford University Press.

Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews. Neuroscience, 6*(2), 159–164.

Racine, E., Bar-Ilan, O., & Illes, J. (2006). Brain imaging: A decade of coverage in the print media. *Science Communication, 28*(1), 122–142.

Racine, E., Waldman, S., Palmour, N., Risse, D., & Illes, J. (2007a). Currents of hope: Neurostimulation techniques in US and UK print media. *Cambridge Quarterly of Healthcare Ethics, 16*(3), 314–318.

Racine, E., Van der Loos, H. Z. A., & Illes, J. (2007b). Internet marketing of neuroproducts: New practices and healthcare policy challenges. *Cambridge Quarterly of Healthcare Ethics, 16*(2), 181–194.

Racine, E., Amaram, R., Seidler, M., Karczewska, M., & Illes, J. (2008). Media coverage of the persistent vegetative state and end-of-life decision-making. *Neurology, 71*(13), 1027–1032.

Racine, E., Waldman, S., Rosenberg, J., & Illes, J. (2010). Contemporary neuroscience in the media. *Social Science and Medicine, 71*(4), 725–733.

Rose, S. P. R. (2003). How to (or not to) communicate science. *Biochemical Society Transactions, 31*(2), 307–312.

Safire, W. (2002). Neuroethics belongs in public eye . *Dayton Daily News*, p. 13A.

Sellars, W. (1963). *Science, perception, and reality*. New York: Humanities Press.

Striano, P., Bifulco, F., & Servillo, G. (2009). The saga of Eluana Englaro: Another tragedy feeding the media. *Intensive Care Medicine, 35*(6), 1129–1131.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49–54.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477.

Wijdicks, E. F., & Wijdicks, C. A. (2006a). The portrayal of coma in contemporary motion pictures. *Neurology, 66*(9), 1300–1303.

Wijdicks, E. F., & Wijdicks, M. F. (2006b). Coverage of coma in headlines of US newspapers from 2001 through 2005. *Mayo Clinic Proceedings, 81*(10), 1332–1336.

Williams, S. J., Seale, C., Boden, S., Lowe, P., & Steinberg, D. L. (2008). Waking up to sleepiness: Modafinil, the media and the pharmaceuticalisation of everyday/night life. *Sociology of Health and Illness, 30*(6), 839–855.

Popular Media and Bioethics Scholarship: Sharing Responsibility for Portrayals of Cognitive Enhancement with Prescription Medications

93

Cynthia Forlini, Brad Partridge, Jayne Lucke, and Eric Racine

Contents

C. Forlini (✉)
Institut de recherches cliniques de Montréal (IRCM), Neuroethics Research Unit, Montréal, QC, Canada

UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia
e-mail: c.forlini@uq.edu.au

B. Partridge • J. Lucke
University of Queensland Centre for Clinical Research, The University of Queensland, Brisbane, QLD, Australia
e-mail: b.partridge@sph.uq.edu.au; j.lucke@uq.edu.au

E. Racine
Neuroethics Research Unit, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

Department of Medicine and Department of Social and Preventive Medicine, Université de Montréal, Montréal, QC, Canada

Departments of Neurology and Neurosurgery, Experimental Medicine & Biomedical Ethics Unit, McGill University, Montréal, QC, Canada
e-mail: eric.racine@ircm.qc.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_83,

Abstract

The use of prescription medications for the cognitive enhancement of healthy individuals has been a captivating and contentious subject in the media and in the bioethics literature. This chapter provides an overview of the roles the media have played in reporting features of the nonmedical use of prescription medications by healthy individuals for enhancement and the influence the media have exerted on academic discourses. We note the broadening of media coverage from trend reporting to taking on research questions of the prevalence and efficacy of cognitive enhancement. We also reflect on the emergent contribution of messages via social media in online discussions about cognitive enhancement. Messages from traditional and social media are often based on anecdotal evidence and may perpetuate misrepresentations of cognitive enhancement especially when referenced by academic researchers. Such misrepresentations may harbor underlying normative messages that can influence the behavior of individuals. The different forms of media and academic bioethics share a responsibility to portray cognitive enhancement in a light that reflects current evidence and does not encourage the nonmedical use of medications based on anecdotal experiences.

Introduction

The use of prescription medications by healthy individuals for cognitive enhancement has been a captivating subject in the media and in the bioethics literature. The scientific and ethical underpinnings of cognitive enhancement have polarized the perspectives of contributors to the bioethics discussion, leaving ample room for debates on scientific evidence and ethical principles (Forlini and Racine 2013). Over the last 10 years, a significant number of newspaper articles and special reports have captured the attention of readers as well as other audiences via television and radio reports or discussion over social media (Forlini and Racine 2009b; Partridge et al. 2011; Wade et al. under review; Williams et al. 2008). Cognitive enhancement is often *reported* as a phenomenon growing in prevalence and complexity. Through various forms of media coverage, journalists are seeking to relay information about and expose the underpinnings of a new or contentious phenomenon (Seale 2003). In some instances, the search for answers in media reports has blurred the boundaries between investigative journalism and research. A few studies have examined the specific features of media coverage on cognitive enhancement (Forlini and Racine 2009b; Partridge et al. 2011; Wade et al. under review; Williams et al. 2008). This chapter aims to critically examine some of the features of media coverage on cognitive enhancement in order to reflect and comment on the role it has played upstream in the bioethics debate and downstream for public understanding.

How Has Cognitive Enhancement Been Represented in the Media?

Within the following section of this chapter, we identify and discuss three distinct features of media coverage on cognitive enhancement. These features are related to the way journalists engage with the issues in the cognitive enhancement debate. In this respect, we examine: (i) the traditional reporting of a trend, (ii) the transition of the journalist into a "researcher" to fill gaps in knowledge, and (iii) the influence of social media that is shifting the ways in which information and experiences are disseminated and consumed.

Media Coverage Reporting a Trend

In reporting a trend, the journalist relies on already published material or other observations. Early reports on what is now more commonly referred to as cognitive enhancement were based on prevalence studies carried out on US college campuses in the early 2000s (Babcock and Byrne 2000; Graff Low and Gendaszek 2002; McCabe et al. 2004; Teter et al. 2003). These media reports addressed the broader topic of stimulant abuse and sometimes included related aspects of the medical profession and campus culture. Of particular interest in examining trend reporting is the newspaper coverage from the UK on this topic. These newspaper articles characterized the nonmedical use of stimulants for enhancement purposes as a distinctively North American phenomenon that would eventually catch on in the UK as well (Forlini and Racine 2009b). Interestingly, at the time, no study existed to confirm that stimulant use for enhancement was not already occurring in the UK and prevalence figures in the UK generated by academic research have still not been published. The trend reporting proved to be speculative by portraying cognitive enhancement as an epidemic based on data from another cultural context. The US prevalence numbers have since been questioned as to their robustness in supporting the portrayal of cognitive enhancement as a widespread or growing phenomenon (Lucke et al. 2010). In this regard, journalists relied on academic sources whose interpretations were already enthusiastic. Furthermore, in reporting cognitive enhancement as a trend spreading to the UK, journalists ran the risk of potentially accelerating the spread of nonmedical stimulant use. A study conducted by Dasgupta et al. (2009) demonstrated a temporal association between reports in the news media on prescription opioid use and overdose mortality. With news media reports on opioid abuse consistently preceding peaks of poisoning mortality, the article begs the question of whether the news media might have fueled an epidemic of opioid overdoses (Dasgupta et al. 2009). At this point, any influence that the media may have had on prevalence of stimulant use by university students in the UK cannot be verified. The perceived spread of a trend such as cognitive enhancement could have been an informative avenue for academic research into either the actual prevalence of stimulant use among students in the UK or perhaps

a study of whether this type of trend would be troubling to the public in that region. At the time, the trend reports from the UK media might have served as cues for valuable research to characterize an oncoming phenomenon.

Another feature of trend reporting has been the use of anecdotal reports to introduce the topic of nonmedical stimulant use by students. The experience of university students who used stimulants with the intention of enhancing their academic performance has acted as a cornerstone of trend reporting. Reports based on anecdotal evidence provide an experiential anchor to the phenomenon of cognitive enhancement. Partridge et al. (2011) showed that anecdotal evidence was the basis of 15 % of newspaper articles they sampled. Anecdotal evidence of academics using modafinil as a cognitive enhancer also fueled the premise of a commentary published in *Nature* (Sahakian and Morein-Zamir 2007). Using anecdotal evidence in media reports can be problematic on two fronts. First, comments from members of the stakeholder groups in a qualitative study showed that some important pieces of information were deemed to be missing from newspaper articles (i.e., scientific data, risks and benefits, prevalence) (Forlini and Racine 2012). Given the space constraints of some media formats, it would be inappropriate if anecdotal evidence is privileged over academic references. Second, anecdotal evidence explaining the particular experience of users often reveals some of the specific details of how people go about using stimulants and other medications for enhancement. For example, media articles may explain how and where students are procuring medications; imply effective dosages for cognitive enhancement; report the (side) effects that can be expected; and convey resounding praise for enhancers by quoting users. In addition to raising awareness about the cognitive enhancement trend, media reports including anecdotal evidence may inadvertently provide crucial details which would enable readers to do the same. These problems are not confined to media reporting. Academic research on any one of these aspects of prevalence could yield and disseminate the same type of information. This situation highlights the shared responsibility of journalists and academics in walking the fine line between informing and encouraging when reporting on or investigating cognitive enhancement.

A Journalistic-Empirical Approach to Media Coverage

The second feature of media reports on cognitive enhancement concerns the convergence of journalism and research. Instead of relaying information about a trend, some journalists have undertaken investigations that further characterize elements of cognitive enhancement that remain elusive. The prevalence of cognitive enhancement and the efficacy of medications in healthy individuals are of particular interest for these types of reports. In this way, journalists are gaining firsthand knowledge of prevalence in their environment or verifying the anecdotal reports of the effects of stimulants and other enhancing substances on their personal activities. These forms of data are often more accessible to lay audiences than publications in academic journals and would speak to local contexts. However, the results of these journalistic-empirical investigations can lack in the expertise, ethics oversight, and rigor that academic research abides by.

Journalists Researching Prevalence

Much of the media and academic discourse remains centered on the use of stimulants by university students where prevalence studies state that between 2 % and 11 % are using cognitive enhancement to improve academic performance (Franke et al. 2011; Racine and Forlini 2010). However, within the academic debate, these prevalence figures have been subject to much criticism on the basis of their methodology and their interpretation of these findings (Hall and Lucke 2010; Lucke et al. 2010). One of the major limitations of current prevalence studies on the nonmedical use of prescription stimulants for recreation and enhancement is their ability to serve as indicators of general prevalence. They are focused mainly on students in certain regions of North America (Smith and Farah 2011). Only within the last few years has the evidence base grown to become more internationally representative (Castaldi et al. 2012; Franke et al. 2011; Partridge et al. 2012). Students, however, are still the focus of the majority of these studies. Curiosity about where, and among which other groups, cognitive enhancement occurs has spurred some interest in carrying out informal studies. One often cited example is a poll conducted by the scientific journal, *Nature* (Maher 2008). Based on responses of 1,400 respondents from 60 countries, the poll revealed that one in five respondents had used medications nonmedically to increase focus, concentration and memory. Another and more specific example of an unofficial, yet published, prevalence study is a survey of Cambridge University students conducted by the university's newspaper, *Varsity*. Of the 100 students surveyed, 1 in 10 admitted to using medications for cognitive enhancement (Lennard 2009). The results of a similar survey conducted by *Newsnight* and the *New Scientist* magazine were published on the BBCNews website (Watts 2011). Of the 761 respondents 38 % said they had taken a drug for cognitive enhancement. The vastly different results of the unofficial prevalence studies could be a reflection of actual variations in prevalence between regions and groups. However, the results cannot be compared or verified because the methods are not available for review nor have the publications undergone peer review.

Reporting and gathering of prevalence data on cognitive enhancement is creating what may be problematic interplay between the various forms of media and bioethics scholarship. Partridge et al. (2011) found that the media articles in their sample most often reported cognitive enhancement as being common or increasing in prevalence. The authors observed that the media often misinterpreted prevalence data. What is perhaps most concerning about this finding is that the misinterpretations of prevalence from media articles have appeared as references in scholarly papers supporting "a trend toward growing use of prescriptions stimulants by healthy individuals for the purpose of enhancing school or work performance" (Smith and Farah 2011 p.717). Informal studies of prevalence conducted by journals and newspapers have also been cited in academic circles as an indication of increasing or high prevalence. Despite a lack of demographic data of the sample, the results from the *Nature* poll have been used by scholars to support claims that cognitive enhancement is not only widespread but also in demand (Greely et al. 2008). The *Varsity* figures have also been cited in the academic literature as sources supporting increasing use of medications for cognitive enhancement

(Mohamed and Sahakian 2011). The uptake of media reports indicating increasing prevalence in bioethics discourse creates a situation where it is difficult to ascertain whether the impression of increasing prevalence originated in the media or the bioethics literature. Both discourses seem to be contributing to the inflation of the perceived high prevalence of cognitive enhancement among university students (Lucke et al. 2011). For now, these types of journalistic-empirical studies are giving credibility to unverified prevalence rates, while the academic empirical studies and their interpretation remain contentious.

Journalists Researching the Efficacy of Medications on Healthy Individuals

Another key element of media reports on cognitive enhancement is the evidence supporting the efficacy of purported cognitive enhancers on healthy individuals. References in the media to controlled trials of efficacy appear inconsistently. Consequently, readers may feel inadequately informed regarding the scientific basis of the benefits and risks mentioned in the report (Forlini and Racine 2012). Existing evidence from clinical trials does not currently support the use of prescription medications in healthy individuals for enhancement (Repantis et al. 2008, 2010a, b). Only a few studies have demonstrated that stimulants produce a modest benefit for lower-performing individuals (Ilieva et al. 2013). Yet, anecdotal reports of efficacy are pervasive in media reports (Schwarz 2012; Talbot 2009). In fact, a study of media coverage on the nonmedical use of methylphenidate in university students showed that the media reported the benefits of cognitive enhancement more often than the risks (Forlini and Racine 2009b). The concern remains that media reports emphasizing the benefits of cognitive enhancement and describing anecdotal efficacy would encourage the nonmedical use of prescription medications for enhancement purposes (Forlini and Racine 2012; Partridge et al. 2011). The bioethics literature has also been the target of criticism for over-enthusiasm about cognitive enhancement (e.g., Bostrom and Sandberg 2009; Greely et al. 2008) despite the lack of efficacy data to support the use of medications in healthy individuals stressed by some (Racine and Forlini 2009). In addition to prevalence, efficacy of medications has provided an element for both the media and those contributing to bioethics scholarship to speculate on. As a result, they both risk portraying the efficacy of purported enhancers in an enthusiastic light.

The gap in knowledge about efficacy has driven some journalists to seek a better, and often more personal, grasp on the effects of medications on healthy individuals. On the one hand, many journalists have approached individuals who do engage in cognitive enhancement to explain the effects they experienced. On the other hand, some journalists have opted to explore the subjective effects themselves. By self-experimenting, the journalist acts as both researcher and research subject in order to gain information about yet another aspect of cognitive enhancement whose evidence base is contentious. These journalists set the plan for what they will experiment with and when (researcher) but are also subjectively reporting the results of their informal study (research subject). The self-experimentation of

journalists, however, establishes a basis for anecdotal efficacy evidence that can be cited with the same deceptive authority similar to unofficial surveys.

There are several notable examples of the journalist-researcher. In one article, Hyland (2013) orders modafinil from a website upon hearing of a friend's positive experience using the medication to be more productive. Her article focused on how modafinil could help her cope with the fatigue caused by multiple sclerosis, a condition that is not part of modafinil's indicated uses. She clearly states that "if it worked, I'd take my anecdotal evidence of the drug's effects to both my neurologist and my GP" (Hyland 2013). Thus, this experiment was done without any type of medical supervision. After experiencing positive results, the author carried on using modafinil she ordered from the Internet until her neurologist was able to issue a legitimate prescription (following a refusal from her primary care physician and resistance from healthcare services).

The author of a 2012 feature in *Rolling Stone* Magazine also experimented with modafinil and other marketed nootropics available on the Internet (McMillen 2012). McMillen reported experiencing positive effects from modafinil and other substances he experimented with. This reporter also diligently documented the negative side effects that he experienced.

A third example comes from a BBCNews reporter. In contrast to Hyland and McMillen, Watts sought the supervision of a neurologist. The physician carried out cognitive testing to evaluate the reporter's performance while using modafinil on one day as compared to a placebo on another day. Though the results showed that Watts performed better using modafinil, when asked to guess which substance she had taken on which day, she was unable to identify the day she took modafinil (Watts 2011).

Similar to the trend reporting, articles based on journalist self-experimentation provide a how-to guide to would-be users of medications for cognitive enhancement. However, the first person narrative of self-experimentation reports contains some added incentives for would-be users. The means employed to obtain substances for experimentation are usually detailed and, most of all, successful. Both Hyland and McMillen demonstrated how easily one could procure medications online. Examples of how drastically the substances impacted the work and productivity of the journalist-researcher that tip the balance in favor of the positive effects might empower readers into following the self-experimentation of journalists. Consequently, these types of reports could undermine measures aimed at reducing the nonmedical use of medications, regardless of the real-world efficacy of cognitive enhancers.

There are two important and intimately linked ethical issues that arise specifically from reports based on self-experimentation. The first issue is the safety of experimenting with medications for unregulated purposes. The second is the legal aspects generally related to the nonmedical use of medications. Watts was supervised by a neurologist who verified that modafinil would pose no risk to the journalist. There were also medically trained staff nearby during the experiment (Watts 2011). However, according to Downie et al. the physician would be liable for any unforeseen harm that could have befallen the journalist during the cognitive enhancement experiment (Downie et al. 2010). McMillen's article does acknowledge the risk of taking the prescription medications and other nootropics with which

he experiments. A disclaimer at the end of the article warned that, "At no point should any of the products mentioned in this article be ingested without first consulting a health professional" (McMillen 2012). Hyland's article features no such disclaimer. This disclaimer is problematic on two fronts. Though it acknowledges risks, the disclaimer sends a contradictory message and sets a poor example for readers when the journalist had not taken that same precaution. The disclaimer seems to protect the author and publisher from liability more than the safety of the individual who may decide to follow the example of the journalist in trying medications or substances for enhancement purposes. Hyland and McMillen's experiments evoke legal aspects related to the purchase and possession of prescription medications without a prescription. Both these journalists purchased modafinil over the Internet. Aside from not being able to verify the authenticity of the medications they were receiving and ingesting, it is technically illegal for them to possess these drugs (Downie et al. 2010). Even if there is an acknowledgement that these drugs could be harmful and are obtained from illicit sources, there still seems to be tolerance for experimentation with suggested cognitive enhancers for journalistic endeavors in spite of national public health agencies attempting to curtail nonmedical uses of prescription drugs. Different frameworks that cut across disciplinary boundaries have been used to discuss cognitive enhancement (Outram 2012; Racine and Forlini 2010). Previous work showed that the media adhered largely to a framework that portrayed cognitive enhancement as a lifestyle choice (Forlini and Racine 2009b). However, since this work was published, some recent reports have adopted frameworks that reflect the ethical discussions on media prescription practices (Schwarz 2012) and investigate what the effects and benefits of purported cognitive-enhancing medications for healthy individuals might be (McMillen 2012).

Reports over Social Media: The Public Becomes the Journalist

In the previous sections of this chapter, we mentioned two sources of information on cognitive enhancement: academic research and the popular media. In the context of cognitive enhancement, social media warrant attention as another source of public information for those who engage with the range of social media available, many of whom are young people. Social media include networking sites such as Facebook and Twitter, and personal blogs. With social media, the research and writing is not only being done by professionals (i.e., researchers or journalists) but also by anyone with an opinion or an experience to share. In this way, members of the public are consuming and *creating* messages about cognitive enhancement. Social media harness the power of anecdotal reports just as the traditional media do, and academic papers have been inclined to do, in the context of cognitive enhancement (Sahakian and Morein-Zamir 2007). The unique power of social media is that the sources of the anecdotes are friends, acquaintances, and others that one can *choose* to follow.

There is a cacophony of chatter about cognitive enhancement over social media. Internet searches can reveal declarations of need for cognitive enhancers, requests for sources to obtain cognitive enhancers illicitly, and discussion of the effects of medications. There are blogs explaining the effects of various substances including

the legal implications (Anonymous 2012) and YouTube videos conducting informal efficacy trials ("Smart Drugs and Memory Enhancers," 2012). Talk of cognitive enhancement using prescription medications over social media is largely uncharted territory for academic research. Hanson et al. conducted a highly relevant study that monitored tweets containing the term "Adderall" and found 213,633 tweets from 132,099 unique user accounts over 6 months in the USA (Hanson et al. 2013). The authors reported that:

> The number of Adderall tweets peaked during traditional college and university final exam periods. Rates of Adderall tweeters were highest among college and university clusters in the northeast and south regions of the United States. 27,473 (12.9 %) mentioned an alternative motive (e.g., study aid) in the same tweet" (Hanson et al. 2013).

Further examination of the cognitive enhancement content of social media would yield insightful information on how cognitive enhancement is being portrayed in informal and often very short messages. Hanson et al. conclude that the content of social media can contribute to the formation of normative behaviors regarding the use of stimulants and other medications for enhancement purposes. A previous qualitative study of different stakeholder groups similarly demonstrated how perceived acceptability or necessity of a practice like cognitive enhancement can have an impact upon individual behaviors (Forlini and Racine 2009a). Information over social media is disseminated at an astonishingly rapid rate. With the potential to "go viral," the impact of blogs, tweets, and status updates on normative behaviors can be more influential than traditional forms of media.

Bioethics Literature and the Media: Sharing the Responsibility to Discuss and Reflect on Cognitive Enhancement

The media have contributed to various facets of the cognitive enhancement debate and bioethics scholarship. The subject of cognitive enhancement has permeated popular culture, sparking various representations of what it means to enhance cognition and what the consequences may be (McKenna 2011a, b). Some of the media's interest for cognitive enhancement might be explained by the multifaceted nature of the phenomenon. These types of contributions have broadened the scope of media presence in the cognitive enhancement debate. However, distinguishing the anecdotal evidence from scientific evidence in media reports on cognitive enhancement may not always be obvious for audiences. The experiences of reputable journalists and social media contacts can be more accessible than academic research. Furthermore, testimonials from these sources carry weight with readers and followers. Pop culture representations have also been shown to reinforce stereotypes of cognitive enhancement "as an admirable attempt to succeed in a troubled system" (McKenna 2011a p. 92). In a joint guide for journalists, the UK Drug Policy Commission and the Society for Editors acknowledge the policy and social implications of media reports on illicit drug use. The guide cautions against relying on anecdotal or uncertain evidence about substances and employing language that reinforces drug-related stigma (Society of

Editors and UK Drug Policy Commission 2012). Piercing through the dissemination of anecdotal evidence and reinforcement of stereotypes in popular culture is a complicated task for researchers who are trying to communicate nuanced messages about cognitive enhancement, especially regarding its prevalence and efficacy. Journalists face an equal challenge in discussing aspects of a socially relevant phenomenon where evidence is absent or disputed.

In seeking and reporting anecdotal information, journalists balance their responsibility to convey information about a phenomenon against the potential to encourage the nonmedical use of medications for cognitive enhancement. Stakeholders have been shown to value the information in media reports while being concerned with media coverage that encourages further use (Forlini and Racine 2012). Encouragement, whether intentional or collateral, might come through demonstrating the ease with which medications can be obtained without a prescription, describing positive or life-changing effects of enhancers, and describing perceptions of acceptability of cognitive enhancement among certain groups. Concerns about media reporting of drugs and drug addiction prompted the Australian Press Council to release practice guidelines (Australian Press Council 2001). In essence, these guidelines urge journalists to guard against reporting of details that would facilitate or encourage self-experimentation relating to illicit drug use and abuse. Given the circumstances of how prescription medications for cognitive enhancement are obtained, the potential negative effects and lack of medical supervision, following the guidelines of the Australian Press Council, would be a relevant precaution for journalists to take when reporting on cognitive enhancement, particularly when it involves illegal activity.

Various forms of media are often blamed for misinterpreting, misrepresenting, or hyping scientific evidence which, ultimately, skews public understanding (Seale 2003). The goal of this chapter was not to criticize media reports on cognitive enhancement per se but rather to characterize the role popular media have played in the debate (including its academic counterpart) and the types of information it has contributed. Moreover, we sought to raise awareness of the contribution of academic bioethics literature in subscribing to anecdotal evidence and elements of hype. In a recent report, the Nuffield Council on Bioethics acknowledged the contribution of researchers and media alike in the "spiral of hype" for novel neurotechnologies (Nuffield Council on Bioethics 2013 p. 17). The Council recommended that "Researchers, press officers and journalists . . . reflect on how their representations might contribute to hype, and exercise caution when describing the possible applications of a technology" (Nuffield Council on Bioethics 2013 p. 17). The very term ("cognitive enhancement") used in bioethics discourse has been critiqued for the assumptions and expectations for benefit that it creates (Outram 2012; Racine and Forlini 2009). In this sense, the media cannot be faulted for following the cues of bioethicists. Responsibility for messages about cognitive enhancement is shared between the media and those contributing to bioethics scholarship. Many groups have suggested ways to improve science communication (Bubela et al. 2009; Illes et al. 2010), and recent projects have ventured into more interventional approaches to match scientific expertise with journalistic endeavors (Callaway 2013). Improving scientific communication is not only about improving the scientific

literacy of members of the media, but also making researchers more media-literate. In addition, others have recommended that bioethics scholarship adopts more explicitly methodological approaches (e.g., acknowledging assumptions, validating assumptions, adopting broader frameworks) to prevent hasty leaps of faith (Racine et al. under review). An outstanding question, however, remains whether and how social media will be influenced by changes in scholarly papers or media reports. In Box 1, we make some suggestions regarding avenues for future research that could help understand and respond to media messages on cognitive enhancement in addition to identifying the aspects of cognitive enhancement research which are vulnerable to misinterpretation.

Box 1: Recommendations for Academic Research Examining the Content and Impact of Media Reports on Cognitive Enhancement

Implications of Conventional and Social Media for Academic Research

- Measurement of the impact of media coverage of enhancement controversy on academic scholarship
- Demonstration of the impact of expectations of public visibility of bioethics scholarship on attitudes defended by bioethics scholars in academic literature and in the media
- Recommendations for proper use of levels of evidence in the media and academic scholarship
- Characterization of the impact of various pressures for "translation" of academic research on media portrayal of cognitive enhancement
- Understanding the relationship between journalists and academic researchers

Content and Impact of Conventional and Social Media

- Further characterization of the content of international conventional media as well as social media
- Understanding patterns of social media use for disseminating messages on cognitive enhancement
- Comparison of conventional versus social media on attitudes and behaviors in specific stakeholder groups
- Understanding diverging practices for media content on illicit drugs versus nonmedical use of prescription drugs

Conclusion

Media reports and academic publications both strive to collect and disseminate information. However, they differ with respect to their audiences, methods, and

formats, which, in turn, relate to the distinction between the tasks of journalists and those of researchers. It is important to keep these distinct roles in mind when studying media portrayals of phenomena like cognitive enhancement that beg both scientific as well as social and ethical questions. Collaborative projects between researchers and journalists are one way of improving communication about evidence, practices, and attitudes related to cognitive enhancement (Lucke, Partridge, & Hall 2012). However, an essential part of studying cognitive enhancement from ethical and journalistic perspectives is an awareness of the shared commitment and responsibility in accurately representing evidence and practices.

Cross-References

▶ Neuroenhancement
▶ Reflections on Neuroenhancement
▶ Smart Drugs: Ethical Issues

References

Anonymous. Gwern.net. Retrieved October 18, 2012, from http://www.gwern.net/Nootropics#results.

Australian Press Council (2001) Reporting guidelines: Drugs and drug addiction general press release No. 246 (ii). http://www.presscouncil.org.au/document-search/guideline-drugs-and-drug-addiction/?LocatorGroupID=662&LocatorFormID=677&FromSearch=1. Accessed 18 October 2013.

Babcock, Q., & Byrne, T. (2000). Student perceptions of methylphenidate abuse at a public liberal arts college. *Journal of American College Health, 49*(3), 143–145.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311–341.

Bubela, T., Nisbet, M. C., Borchelt, R., Brunger, F., Critchley, C., Einsiedel, E., . . . Caulfield, T. (2009). Science communication reconsidered. *Nature Biotechnology, 27*(6), 514–518.

Callaway, E. (2013). Science media: Centre of attention. *Nature, 499*(7457), 142–144.

Castaldi, S., Gelatti, U., Orizio, G., Hartung, U., Moreno-Londono, A. M., Nobile, M., & Schulz, P. J. (2012). Use of cognitive enhancement medication among northern Italian university students. *Journal of Addiction Medicine, 6*(2), 112–117.

Dasgupta, N., Mandl, K. D., & Brownstein, J. S. (2009). Breaking the news or fueling the epidemic? Temporal association between news media report volume and opioid-related mortality. *PLoS ONE, 4*(11), e7758.

Downie, J., Outram, S., & Campbell, F. (2010). Caveat emptor, venditor, et praescribor: Legal liability associated with methyplenidate hydrochloride (MPH) use by postsecondary students. *Health Law Journal, 18*, 51–71.

Forlini, C., & Racine, E. (2009a). Autonomy and coercion in academic “cognitive enhancement” using methylphenidate: Perspectives of a pragmatic study of key stakeholders. *Neuroethics, 2*(3), 163–177.

Forlini, C., & Racine, E. (2009b). Disagreements with implications: Diverging discourses on the ethics of non-medical use of methylphenidate for performance enhancement. *BMC Medical Ethics, 10*, 9.

Forlini, C., & Racine, E. (2012). Stakeholder perspectives and reactions to “academic” cognitive enhancement: Unsuspected meaning of ambivalence and analogies. *Public Understanding of Science, 21*(5), 606–625.

Forlini, C., & Racine, E. (2013). Does the cognitive enhancement debate call for a renewal of the deliberative role of bioethics? In E. Hildt & A. Franke (Eds.), *Cognitive enhancement: An interdisciplinary perspective* (pp. 173–186). New York: Springer.

Franke, A. G., Bonertz, C., Christmann, M., Huss, M., Fellgiebel, A., Hildt, E., & Lieb, K. (2011). Non-medical use of prescription stimulants and illicit use of stimulants for cognitive enhancement in pupils and students in Germany. *Pharmacopsychiatry, 44*(2), 60–66.

Graff Low, K., & Gendaszek, A. E. (2002). Illicit use of psychostimulants among college students: A preliminary study Psychology. *Health and Medicine, 7*(3), 283–287.

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456*(7224), 702–705.

Hall, W. D., & Lucke, J. C. (2010). The enhancement use of neuropharmaceuticals: More specticism and caution needed. *Addiction, 105*(12), 2041–2043.

Hanson, C. L., Burton, S. H., Giraud-Carrier, C., West, J. H., Barnes, M. D., & Hansen, B. (2013). Tweaking and tweeting: Exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research, 15*(4), e62.

Hyland, M. (2013). The drugs do work: My life on brain enhancers, *The Guardian*. Retrieved from http://www.guardian.co.uk/lifeandstyle/2013/may/03/brain-enhancing-drugs-mj-hyland. Accessed 26 July 2013.

Ilieva, I., Boland, J., & Farah, M. J. (2013). Objective and subjective cognitive enhancing effects of mixed amphetamine salts in healthy people. *Neuropharmacology, 64*(1), 496–505.

Illes, J., Moser, M. A., McCormick, J. B., Racine, E., Blakeslee, S., Caplan, A., . . . Weiss, S. (2010). Neurotalk: Improving the communication of neuroscience research. *Nature Reviews. Neuroscience, 11*(1), 61–69.

Lennard, N. (2009, 6 March 2009). One in ten takes drugs to study. *Varsity*. http://www.varsity.co.uk/news/1307. Accessed 18 October 2013.

Lucke, J., Partridge, B., & Hall, W. (2012). Ritalin rising? Let's be smarter about 'smart drugs'. *The Conversation*. http://theconversation.com/ritalin-rising-lets-be-smarter-about-smart-drugs-8398. Accessed 18 October 2013.

Lucke, J., Bell, S., Partridge, B., & Hall, W. (2010). Weak evidence for large claims contribute to the phantom debate. *BioSocieties, 5*(4), 482–483.

Lucke, J., Bell, S., Partridge, B., & Hall, W. D. (2011). Deflating the neuroenhancement bubble. *AJOB Neuroscience, 2*(4), 38–43.

Maher, B. (2008). Poll results: Look who's doping. *Nature, 452*(7188), 674–675.

McCabe, S. E., Teter, C. J., Boyd, C. J., & Guthrie, S. K. (2004). Prevalence and correlates of illicit methylphenidate use among 8th, 10th, and 12th grade students in the United States, 2001. *Journal of Adolescent Health, 35*(6), 501–504.

McKenna, S. A. (2011a). Maintaining class, producing gender: Enhancement discourses about amphetamine in entertainment media. *International Journal of Drug Policy, 22*(6), 455–462.

McKenna, S. A. (2011b). Reproducing hegemony: The culture of enhancement and discourses on amphetamines in popular fiction. *Culture, Medicine and Psychiatry, 35*(1), 90–97.

McMillen, A. (2012). Building a better brain: Wired on nootropics. *Rolling Stone Australia*, pp. 78–83.

Mohamed, A. D., & Sahakian, B. J. (2011). The ethics of elective psychopharmacology. *International Journal of Neuropsychopharmacology, 15*(4), 559–571.

Nuffield Council on Bioethics. (2013). Novel neurotechnologies: intervening in the brain. London. http://www.nuffieldbioethics.org/sites/default/files/Novel_neurotechnologies_report_PDF_web_0.pdf. Accessed 26 July 2013.

Outram, S. M. (2012). Ethical considerations in the framing of the cognitive enhancement debate. *Neuroethics, 5*(2), 173–184.

Partridge, B., Bell, S., Lucke, J., Yeates, S., & Hall, W. (2011). Smart drugs "As Common As Coffee": Media hype about neuroenhancement. *PLoS ONE, 6*(11), e28416.

Partridge, B., Lucke, J., & Hall, W. (2012). A comparison of attitudes toward cognitive enhancement and legalized doping in sport in a community sample of Australian adults. *AJOB Primary Research, 3*(4), 81–86.
Racine, E., & Forlini, C. (2009). Expectations regarding cognitive enhancement create substantial challenges. *Journal of Medical Ethics, 35*(8), 469–470.
Racine, E., & Forlini, C. (2010). Cognitive enhancement, lifestyle choice or misuse of prescription drugs? Ethical blindspots in current debates. *Neuroethics, 3*(1), 1–14.
Racine, E., Martin Rubio, T., Chandler, J., Forlini, C., & Lucke, J. (under review). The value and pitfalls of speculation about science and technology in bioethics: The case of cognitive enhancement.
Repantis, D., Schlattmann, P., Lainsey, O., & Heuser, I. (2008). Antidepressants for neuroenhancement in healthy individuals: A systematic review. *Poiesis & Praxis, 6*(3–4), 139–174.
Repantis, D., Laisney, O., & Heuser, I. (2010a). Acetylcholinesterase inhibitors and memantine for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 61*(6), 473–481.
Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010b). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research, 62*(3), 187–206.
Sahakian, B., & Morein-Zamir, S. (2007). Professor's little helper. *Nature, 450*(7173), 1157–1159.
Schwarz, A. (2012, June 10, 2012). Risky rise of the good-grade pill, *The New York Times*. Retrieved from http://query.nytimes.com/gst/fullpage.html?res=9F01E2DE1339F933A25755C0A9649D8B63. Accessed 18 October 2013.
Seale, C. (2003). Health and media: An overview. *Sociology of Health and Illness, 25*(6), 513–531.
Smart Drugs and Memory Enhancers. (2011). Retrieved October 29, 2012, from http://www.youtube.com/watch?feature=endscreen&v=edZbbYuGsYo&NR=1
Smith, E. M., & Farah, M. J. (2011). Are prescription stimulants "smart pills"? The epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological Bulletin, 137*(5), 717–741.
Society of Editors, & UK Drug Policy Commission. (2012). Dealing with the stigma of drugs: A guide for journalists. http://www.societyofeditors.co.uk/userfiles/files/DrugsReportingStigma-251012FinalLowRes.pdf
Talbot, M. (2009, April 27, 2009). Brain Gain. *The New Yorker*.
Teter, C. J., McCabe, S. E., Boyd, C. J., & Guthrie, S. K. (2003). Illicit methylphenidate use in an undergraduate student sample: Prevalence and risk factors. *Pharmacotherapy, 23*(5), 609–617.
Wade, L., Forlini, C., & Racine, E. (under review). Generating genius: A critical examination of how an Alzheimer's drug became a "cognitive enhancer".
Watts, S. (2011). Do cognitive-enhancing drugs work? *BBC News*. Retrieved from http://www.bbc.co.uk/news/health-15600900. Accessed 26 July 2013.
Williams, S. J., Seale, C., Boden, S., Lowe, P., & Steinberg, D. L. (2008). Waking up to sleepiness: Modafinil, the media and the pharmaceuticalisation of everyday/night life. *Sociology of Health and Illness, 30*(6), 839–855.

Neuroethics Beyond Traditional Media

94

Chiara Saviane

Contents

Abstract

Brain science is vital to help us understand how the brain works and identify brain disease treatments. Fast technological advances are opening doors to clinical and nonclinical applications which may affect our health as well as different aspects of our society, from education to business and criminal justice. The scope, benefits, and risks of new technologies and therapies are still uncertain, but they raise crucial ethical, social, and legal issues which involve people from all walks of life. Citizens need to acquire the competences to make informed choices and contribute to decision-making processes which may be critical for their life and the society they want to live in. To this aim, several activities have been promoted over the last decade at the European level, thanks to key players such as the European Commission and the European Dana Alliance for the Brain.

This chapter describes in detail some major European initiatives which have used different approaches beyond traditional media to raise public awareness and

C. Saviane
Interdisciplinary Laboratory for Advanced Studies, Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
e-mail: saviane@sissa.it

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_87,

engagement in neuroscience and neuroethics, providing sound information, fostering multidisciplinary debates and participation to policy-making processes. These include European projects, such as Brains in Dialogue, Meeting of Minds and Decide, and many activities promoted by the European Dana Alliance for the Brain. Some impediments still exist for a true dialogue and a real involvement of citizens in decision-making processes; however, some successful approaches, ranging from public events to social media, have been identified and could be developed and investigated further.

Neuroscience and Society

Over the recent years, philosophers, social scientists, and some neuroscientists have started to investigate the ethical, legal, and social implications of brain science. However, these issues concern so many aspects of our life and our society that, with time, the urge of a multidisciplinary approach as well as the importance of public engagement in such discussions have become clear.

Actually the type and quality of the debate on neuroethics differs from country to country (Illes et al. 2005). The United States, followed by Canada, has been the center of neuroethics activity in the twenty-first century. Many events have been held to define the field and also to engage the public, fostering a deeper understanding of the issues as well as a lively participation in the dialogue already from the beginning. In Europe, the approach to public engagement is still changing and not without some difficulties. The Bodmer report and the following establishment of a Committee on the Public Understanding of Science in 1986 marked the beginning of a series of educational activities in the United Kingdom, and then in Europe, to bridge science and society. However, the need of a shift from a top-down approach to a two-way communication in the relationship between science and society became evident only at the beginning of the twenty-first century. In the United Kingdom, a report published in 2000 by the House of Lords Science and Technology Select Committee entitled "Science and society" described the results of a national survey highlighting a "new mood for dialogue" due to the lack of trust in science among British Citizens (http://www.publications.parliament.uk/pa/ld199900/ldselect/ldsctech/38/3801.htm). The report provided different recommendations on how to put this in practice at the national level, focusing on issues that are still crucial today such as risk and uncertainty communication, science education in schools, and the relationship between science and media. The following year a Eurobarometer Survey conducted in fifteen Member States ("Europeans, science and technology," 2001) reported a mixed attitude among European citizens ranging from confidence and high expectations in science to a complete lack of interest or trust accompanied by a concern about the rapid pace of advancement. Thus, in December 2001, the European Commission agreed a "Science and Society Action Plan" with 38 actions for the establishment of a "new partnership" between science and society ("Science and Society Action Plan," 2002). It was clear that, in order to meet the needs of European citizens and regain their support, it

was necessary to foster public awareness, providing understandable and sound information, but it was also important to encourage a true engagement, establishing a dialogue with citizens and giving them the opportunity to express their views in the appropriate bodies.

Since then, the European Commission has supported several initiatives to foster a dialogue between science and society and bring science policy closer to citizens. Some of these were focused on neuroscience like the Brains in Dialogue and the Meeting of Minds projects. Starting from these experiences, this chapter presents some examples of nontraditional approaches and media used over the last decade in Europe to foster public awareness and engagement in neuroscience and neuroethics, in settings that run from real to virtual.

The Brains in Dialogue Project: Brain Science at the Service of European Citizens

The Brains in Dialogue project (BID) was a three-year project supported by the European Commission under the Seventh Framework Programme and coordinated by the International School for Advanced Studies (SISSA) in Trieste, Italy. Focusing on brain imaging, brain devices, and predictive medicine in brain science, the project aimed at fostering public engagement and a multidisciplinary dialogue between key stakeholders, including scientists, clinicians, philosophers, social scientists, lawyers, economists, journalists, science communicators, patients, service users, lay public, etc. With a scientific and communicative mission, BID used different approaches to communicate the state of the art, discuss the expectations, benefits and risks of new technologies and therapies in neuroscience, build constructive discussions on the ethical, legal, and social issues, and test novel "dialogue products" which integrate lay and scientific knowledge. In particular, between 2008 and 2011, BID organized three interdisciplinary workshops and several public events, managed the website *neuromedia corner* (www.neuromedia.eu) and a press office active at a European level (Ramani and Saviane 2010).

The events represented the core of the project, where different stakeholders – including lay citizens – had the opportunity to meet and address specific issues from different perspectives. The workshops were organized as closed meetings with a limited number of participants and ample time for interactions in order to establish an informal environment and facilitate a true dialogue. Speakers and participants from different countries and fields of research were chosen in order to create an interdisciplinary and international group which could also take into account cultural differences. In particular, a great effort was dedicated to the involvement of patients and service users as speakers or participants, in order to bring their views – often missing – in the debate. Science communicators and journalists were also involved to help the discussion as well as to analyze and stress the difficult relationship between scientists and media and the crucial role of media in shaping public opinion.

The workshops involved between 40 and 70 participants including scientists, clinicians, patients, sociologists, lawyers, philosophers, economists, service users, delegates of the European Commission, science communicators, and other experts from up to 25 European and extra-European countries. The meetings comprised sessions of nontechnical talks and facilitated discussions aimed at providing background information on the different aspects of the topic (scientific, clinical, ethical, legal, economic, etc.) and finding a common language. A final session was aimed at testing new dialogue formats, spanning from group activities to discussion games, in order to achieve a lively debate among all participants and encourage them to change perspective for a while. The discussion was always open to the general public with a round table or Café Scientifique organized in collaboration with the local institutes.

In order to reach experts and citizens from several countries, the workshops and public events were organized in different locations across Europe. The first BID workshop – "Brains in dialogue on brain imaging" – took place in Cambridge, UK, and focused on the current and potential applications of brain imaging in psychiatry. The meeting ended with the *Café Scientifique* "Can we read minds?" which was part of the Cambridge Science Festival and the Brain Awareness Week and focused on the scope and limits of brain imaging technologies for mind-reading and their potential use for nonclinical applications such as lie-detection. The second workshop – "Brains in dialogue on genetic testing" – took place in Trieste, Italy, and addressed the state of the art of predictive genetic testing for neurodegenerative diseases. The discussion was opened to the general public with the round table "Health and DNA: my life, my genes". The last workshop – "Brains in dialogue on deep brain stimulation" – was held in Warsaw, Poland. The meeting focused on the potentials and limitations of deep brain stimulation with great attention to the patients' perspective and the role of media. It ended with the *Café Scientifique* "Brain, machine and something in between" which took place as part of the Warsaw Science Festival and addressed the state of the art and the ethical implications of deep brain stimulation and brain machine interfaces.

Among other initiatives, BID also organized the round table "When the final hours come: End of life care, ethics, costs, and the role of the media" as part of the Euroscience Open Forum (ESOF) 2010. It involved clinicians, scientists, and philosophers as well as journalists coming from some of the most important newspapers and magazines in Europe.

The challenges for an appropriate and effective communication to the general public represented a regular topic of discussion throughout the project. The debate among researchers and journalists on the different responsibilities still seems to be alive. Some academics do not see the importance of improving their communication skills or understanding the media logic for an appropriate public communication and a fruitful interdisciplinary dialogue. However, most of BID participants agreed on the need to train scientists in science communication, an old resolution which has not been put into practice yet and is often still a matter of debate. In this regard, BID organized, in collaboration with the European Science

Communication Network (ESConet), a training workshop on neuroscience communication for young researchers from different disciplines interested in the social, ethical, and legal implications of neuroscience.

The discussions started live during the meetings continued virtually on the online hub of BID, the *neuromedia corner*, and more broadly on the web to reach different stakeholders, through the publications of video interviews, interdisciplinary special issues, and articles for national and international newspapers and magazines.

The BID team wrote press releases as well as lay and scientific articles for national newspapers and international journals to reach the general public and academics from different fields. Moreover, the involvement of European science journalists in the workshops generated an international coverage through press articles, radio, and even TV programs. BID edited two special issues on open-access scientific journals with mini-reviews, perspectives, and opinion papers from some of the workshops' participants offering different perspectives: the Research Topic entitled "Emerging issues in brain imaging: a multidisciplinary dialogue" (http://www.frontiersin.org/HumanNeuroscience/researchtopics/emerging_issues_in_brain_imagi/58), published on *Frontiers in Human Neuroscience*, and the Research Topic "The development of deep brain stimulation for neurological and psychiatric disorders: clinical, societal and ethical issues" (http://www.frontiersin.org/IntegrativeNeuroscience/researchtopics/the_development_of_deep_brain_/127), published on *Frontiers in Integrative Neuroscience*. These works represented important opportunities to raise awareness in the neuroscience community about the social, ethical, legal implications of these technologies as well as the patients' perspective and the crucial role of science communication.

The material collected throughout the project was uploaded, and is still available on the website *neuromedia corner* (www.neuromedia.eu), a portal where experts and lay citizens can find original news, scientific content, video interviews, research centers, events, and useful links. The website had a journal-like home page, with news and pictures regularly updated. It contained all outcome material related to BID activities and also provided media operators with useful and understandable news and information. A special section named Viewpoints focused on the social implication of BID scientific areas through a collection of peer-reviewed papers and news items from international scientific and lay journals. In order to increase the visibility of the project and the website and foster the discussion, a Facebook page was also created under the name *neuromediacorner* with links to the project website or to interesting news, pictures of the BID events, and comments. About ten video interviews were recorded during each workshop with key messages from different stakeholders, from academics to clinicians and patients. After editing and approval by the interviewees, these were uploaded on *neuromedia corner* and the *neuromediacorner*'s channel of YouTube (http://www.youtube.com/user/neuromediacorner) to give an overview of the different aspects of the topics addressed. The publication on YouTube made the interviews more accessible for lay surfers and increased their visibility, up to more than 2,000 views per interview, through referral from related videos and specific searches.

On July 2011 BID final conference – "Dialogue to dialogue" – took place in Brussels, Belgium, where three and a half years of initiatives, workshops, publications about neuroscience and its impact on society were presented in a public conference. Despite its success and the enthusiasm of most participants, the BID project revealed that some stakeholders – mainly senior scientists – are still not sufficiently interested or ready to engage in a multidisciplinary dialogue or to open the debate to the general public (Ramani and Saviane 2010). Many of them are so busy and focused on their field of research that they do not see much added value in a better understanding of the real impact of their results on individual's life and society. It is clear that a dialogue between neuroscience and society is an achievement as important as difficult to put into practice and main efforts should be directed toward the younger generations. More experiences like BID, with real opportunities of interdisciplinary and uncommon exchanges and interactions, are still needed. Actually BID was of inspiration for the new European project NERRI – Neuro-Enhancement Responsible Research and Innovation – which involves 18 partners from 11 European countries (www.nerri.eu). NERRI aims to contribute to the introduction of responsible research and innovation in neuroenhancement in the European Research Area and to the shaping of a normative framework underpinning the regulation of neuroenhancement technologies. These will be achieved through mobilization and mutual learning activities engaging scientists, policy-makers, industry, civil society groups, and the wider public.

The Meeting of Minds Project: European Citizens' Deliberation on Brain Science

The Meeting of Minds (http://www.kbs-frb.be/otheractivity.aspx?id=193934&langtype=1033) was a two-year pilot project aimed at involving European citizens in the discussion and assessment of brain science in order to give relevant inputs into European policy-making and foster a wider public debate on this topic. Funded by the European Commission under the Sixth Framework Programme a nd coordinated by the King Baudouin Foundation (Belgium), the project involved 12 institutions from 9 European countries (Belgium, Denmark, France, Germany, Greece, Hungary, Italy, the Netherlands, and the United Kingdom) and a panel of 126 European citizens.

The project was inspired by the increase in participatory technology assessment and foresight in various European national contexts. These approaches had been previously used to discuss scientific and technological issues with experts, policy-makers, citizens, and relevant stakeholders in order to provide policy advice, foster public debate and mutual learning among the different stakeholders. The Meeting of Minds was thought as an attempt to bring this procedure to a European cross-national level, addressing a topic of growing impact for society. The methodology was built on the expertise of the consortium members, which included technology assessment bodies, science museums, academic institutions, and public

foundations, according to a set of criteria which required – among others – an innovative character, a European dimension, a relevance for the public policy, a central role for citizens in the deliberation process, transparency, and accountability.

The project consisted of three National and two European meetings held between 2005 and 2006. It involved 9 national panels of 14 citizens selected in order to guarantee gender balance and a broad range of age and educational background. An advisory board with neuroscientists, psychiatrists, and other stakeholders in the field of brain science was established in each country as guarantee of the quality of the content and the method of the project. The first Introductory National Meetings were held in April–May 2005 with the aim of gathering the panelists and preparing them for the whole process. Six challenging case studies, described in an information brochure, were used to introduce different topics of brain science and the related social and ethical issues: from depression to Alzheimer's disease, from brain imaging to deep brain stimulation, from medicalization to brain enhancement. The opinions, ideas, and concerns raised during the National Meetings were shared during the first European Citizens' Convention in order to get a European perspective. On this occasion, through international group discussions and voting procedures, the panelists identified six broad themes (Regulation and Control; Normalcy vs. Diversity; Public Information and Communication; Pressure from Economic Interests; Equal Access to Treatment; Freedom of Choice) and related sets of questions for the following steps to be taken. Two National Citizens' Assessment Meetings were organized in the following months to give citizens the opportunity to discuss in further depth the themes selected together with scientists, policy-makers, and other stakeholders. The national panels identified key issues within each theme and drafted some recommendations to share in the second, and final, European Citizens' Convention, held in January 2006. On this occasion, the citizens used different dialogue formats – from plenary sessions to carousel sessions and European Cafés – to review the issues, identify the most relevant two for each theme, and then formulate recommendations. At the end of the Convention, a final European Citizens' Assessment report with 37 recommendations on the social, ethical, and legal implications of brain science was presented during a public ceremony at the European Parliament to key policy-makers, scientists, and other stakeholders.

In order to promote the project results and guarantee public visibility of the initiative, the media and the general public were involved at crucial stages of the process at the national and European level. Two European Stakeholders Meetings were also organized during the project to promote the project and create synergies with other initiatives at the European level. Moreover, since January 2006, the different partner organizations, as well as some of the panelists, have been involved in dissemination initiatives at a national and international level. These include presentations in national parliaments, public events, and a broad range of scientific conferences related to science, ethics, and participatory approaches in science and technology.

The Meeting of Minds project still represents a unique example on how to develop and implement a participatory technology assessment process at the European level. As stated in the external evaluation report (Goldschmidt and Renn 2006) “The Project Meeting of Minds accomplished all envisioned objectives – the content related objectives with great success, the procedural objectives with satisfactory success.” In fact, despite some practical problems related to the European Conventions, the project managed to engage citizens in a multi-lingual and multicultural context and allowed them to feel and act as European citizens, providing policy advices in a field of great impact for the individuals and the society. The experiences collected are of great value for the development of this type of participatory procedure and have also highlighted the importance to further engage the citizens in the debate on neuroscience. Unfortunately, the real impact of the results and recommendations at the policy level is unclear, or minor, despite the hopes of the people involved and the efforts made by the organizations. Certainly, too many factors are involved in policy development, especially at the European level.

Discussion Games and Debate Formats

The need to find new approaches to engage the public in science has led to the development of debate formats and discussion games aimed at different targets and different settings.

A successful example is the discussion game PlayDecide (www.playdecide.eu/) created by the European project Decide – Deliberative Citizens’ Debate – in order to understand the potential role of science centers in the democratization of science. The project, started in 2004 and supported by the Scientific Advice and Governance Unit of the European Commission, had three main goals: to raise awareness of the potential of participatory and deliberative consultations, to collect data from debates on controversial issues in science, and to provide a downloadable kit to conduct debates and discussions in science centers (Bandelli and Konijn 2011). It involved museums and science centers, advocacy group associations, and nongovernmental organizations.

PlayDecide is a card game focused on controversial topics in science, including neuroscience. It was inspired by the card game Democs, developed by the New Economics Foundation to help people discuss about politics using visual thinking and participatory methodologies. The game uses three sets of cards (Info, Issue, and Story cards) that provide information and highlight controversial issues to set up discussion groups of up to eight people. The instructions and material needed are available online and downloadable in different languages. After a first briefing phase, the group addresses specific issues to see whether the participants share common views. Finally, the group has to discuss and vote some predefined policies on the topic or even suggest their own. The results can then be uploaded on the website for a comparison across countries. Although a playful activity, PlayDecide is a “serious game” and the clarification of the aim of the discussion is a critical step for its success.

Over the years, users of PlayDecide have started to develop new kits or adapt them to different targets and situations, such as students in the classroom or specific local contexts (Bandelli and Konijn 2011). The consortium realized that the co-development of kits by the users was the added value of the tool. They decided to encourage this by allowing the use of a more relaxed license as well as by providing micro-grants and training for people interested in employing PlayDecide to foster debates and participation.

A kit on "Brain enhancement" is available in 18 languages. Through the different sets of cards, participants learn something and discuss about consciousness, drug resistance, deep brain stimulation, memory-enhancing drugs, and more. More kits could be developed on specific issues of neuroscience and neuroethics and thus become a useful tool for public engagement on those issues. For example, a draft on neuromarketing was prepared by the students of the Master's Course in Science Communication at the International School for Advanced Studies (SISSA) in Trieste, Italy.

A similar initiative which is worth mentioning is the project Citizen Science, ran between 2004 and 2006 and supported by the Wellcome Trust. Citizen Science was set up by the At-Bristol Education team and by the University of Bristol, involving teachers and scientists. It led to the development of different debate formats and settings inspired by games, role play, television chat shows, etc. to engage students in the debate on scientific controversial issues, such as genetic testing, nanotechnology, in vitro fertilization, and alcohol and drug consumption. About 30 events per year were organized in the UK as well as training courses for teachers and science museum staff. The most successful formats are still available online (http://www.at-bristol.co.uk/cz/teachers/Default.htm) and could be of inspiration for the discussion on neuroethical issues.

The BID project, for example, has developed, for one of the workshops, a discussion game built on the model of the "discussion continuum." Divided into groups of eight to nine people, the participants, including experts and patients, were invited to debate about a list of statements covering some critical aspects of deep brain stimulation. Despite the initial skepticism, most of the attendees enjoyed this format of discussion which was a great opportunity for all to express their thoughts and points of view.

The European Dana Alliance for the Brain and the Dana Foundation Website

The European Dana Alliance for the Brain (EDAB: http://www.dana.org/danalliances/edab) plays a crucial role in Europe in fostering public awareness on brain research and its impact on society. Launched in 1997 and modeled on the US-based Dana Alliance for the Brain Initiatives (DABI), EDAB is a nonprofit organization involving more than 200 brain scientists from 29 countries. Entirely supported by the Dana Foundation, EDAB coordinates the Brain Awareness Week in Europe and collaborates with universities, schools, and other organizations to raise public awareness and engagement in brain science.

Started nationwide in 1996 by DABI, the Brain Awareness Week is now a global celebration of the brain for people of all ages (http://dana.org/brainweek/). Every year in March, for a whole week, hundreds of research institutes, schools, and health organizations worldwide contribute to the organization of a rich and diverse program including debates, lectures, experimental workshops, exhibitions, and more. Since its establishment, more than 2,800 partners in 82 countries have participated in the campaign. More and more of these events focus on – or at least touch – the social, ethical, and legal implications of brain science.

Actually, EDAB organizes outreach events on neuroscience and neuroethics throughout the year. A strategic location is the Science Museum's Dana Centre, an adult-only venue in Central London built on purpose to foster public engagement in science in a lively and informal atmosphere through debates, performances, and multimedia facilities. Since the Centre is a collaboration between EDAB, the British Science Association, and the Science Museum, the events cover the whole range of science issues but a focus on neuroscience and neuroethics is predictable since it hosts one of EDAB's offices.

EDAB takes also part in international conferences like the EuroScience Open Forum (ESOF) or the meeting of the Federation of European Neuroscience Societies (FENS). An example is the last William Safire Seminar on Neuroethics held in Barcelona during FENS 2012 and organized in collaboration with the International Neuroethics Society. The seminar entitled "Invading the Brain: What Are the Ethical Issues on Invasive Treatments for Brain Disorders?" touched on the benefits, limits, and ethical implications of new techniques like deep brain stimulation, cell transplantation, and gene therapy. EDAB was also among the eighteen societies contributing to the BNA 2013: Festival of Neuroscience (http://www.bna2013.com/), a unique event organized in occasion of the annual meeting of the British Neuroscience Association and one of the most exciting events in the 2013 neuroscience calendar. Parallel to an intense scientific program that included public and plenary lectures, workshops, symposia, and poster sessions, a rich public engagement program with film screenings, theatrical events, art installations, and a hands-on Street Fair ran to allow the general public to interact with scientists as well as carers, patients, funders, and policy-makers and thus look at neuroscience from different perspectives. In collaboration with the International Neuroethics Society, EDAB was part of a session on "Public Awareness and Societal Impacts" with the workshop "Drugs and Society: The neuroethics of enhancing or erasing memories" focused on the state of the art, benefits, and risks of cognitive-enhancing drugs.

Through the Dana Foundation website (http://www.dana.org), EDAB also provides a wide perspective on the issues related to brain science and its ethical and social implications and offers a connection with a community active worldwide in these fields. EDAB offers a wide range of publications in different European languages, including reports, informative booklets for patients, carers and professionals, and resources for teachers and secondary school students. In particular, the Dana Foundation website includes a full section on neuroethics with original articles, links to news from external sources, information and coverage of events that range from the International Neuroethics Society Annual Meeting in the USA

to the Neuro-Ethics Film Festival in Edinburgh (Label of the 2012 Biomedical Ethics Film Festival). These topics are addressed with a dynamic and interactive approach also through the *Dana Foundation Blog* – which covers news and views on neuroscience, neuroethics, and neuroeducation – as well as on Twitter. The Brain Awareness Week can also be found on Facebook where it has a dedicated page that gives access to an International Calendar and a very useful set of resources including event ideas, planning tips, media and promotional strategies, and educational tools.

Conclusion and Future Directions

Brain science may play a crucial role in different spheres of our life. Not only for our health and access to treatments, but also because of the potential applications in different sectors of society such as education, business, politics, and criminal justice. It is more and more crucial for individuals to acquire the competences to make informed choices and contribute to decision-making processes on issues that may be critical for their life and the society they want to live in.

The debate on neuroethics involves many and different stakeholders and great effort still has to be made to find a common language to tackle these issues. In Europe, as emerged from the BID project, some stakeholders are not even interested in starting a dialogue. However, the experiences described in this chapter show that several initiatives created to raise public awareness and engagement on these issues have grown over the last few years, thanks to the support of key players such as the European Commission and the European Dana Alliance for the Brain.

Several approaches for raising public awareness and engagement beyond traditional media have been proven effective also for neuroscience and neuroethics and could definitely be explored further. From real to virtual, these range from the organization of public events to the use of the web and social media.

Different formats and settings for public events have developed over the years to achieve specific objectives and targets and represent important opportunities to learn and debate also about brain science and its impact on society. Science Festivals are an example among the different public engagement activities not connected to policy outcomes which have rapidly expanded worldwide over the recent years. They combine activities aimed at raising public awareness with others focused on fostering public engagement. A recent study investigated the interests, motivations, and benefits of people attending Science Festivals (Jensen and Buckley 2012). According to this, citizens value, in particular, the possibility of interacting with scientists and the variety of engagement offered for children, adults, and families and report an increased interest and curiosity in new fields of research. More and more often, some of the activities focus on neuroscience and neuroethics. The BID project took part in both Cambridge and Warsaw Science Festival with informal and multidisciplinary events on new applications of neuroscience. Both events turned into lively debates on the ethical and social implications of these technologies involving experts, citizens, and university

students, in particular. The BNA: Festival of Neuroscience took place in London in spring 2013 involving almost 2,000 delegates and more than 5,000 members of the general public (http://www.bna.org.uk/).

Similar opportunities for informal debates are offered on a regular basis by *Café Scientifiques* or Science Cafés (http://www.cafescientifique.org) which represent a forum for discussion in informal venues such as bars, cafés, and bookshops. These events have recently spread worldwide to more than two-hundred (http://www.cafescientifique.org/attachments/article/263/conferencereport.pdf, 2007) and usually address different science topics each time. However, there have also been examples of *Café Scientifiques* focused on neuroscience in Trieste (Italy), in Cardiff (UK), and in Montreal (Canada) as well as on neuroethics in Vancouver (Canada).

Science centers and museums have also become important venues for public awareness and engagement on crucial science and society issues such as climate change, stem cells, or gender gap (Bandelli and Konijn 2011). The Dana Centre is not the only example in Europe of a purpose-built facility to create dialogue opportunities on science, technology, and culture. Science centers and museums are also trying to become key players in the process of participation in the governance of science. The project PlayDecide was developed exactly with this aim and provides a valuable and adaptable tool that could be used further to discuss different issues in neuroethics. It also allows citizens to approach the process of policy-making before they can really play a role in it as the Meeting of Minds project showed that a real involvement in this process is, at present, difficult to achieve, especially at the European level.

The web and social media also represent key tools that should be mobilized further to foster public awareness and engagement on neuroscience and neuroethics. However, there is a central issue related to quality assessment of the information provided and credibility of sources, especially with regard to blogs and social networks. The connection to an established organization, magazine, or research institute may provide a sort of guarantee in this respect. This is the case for inputs by the Dana Foundation and the International Neuroethics Society or for blogs on these topics sponsored by *Nature* (http://blogs.nature.com/actionpotential), *Scientific American* (http://blogs.scientificamerican.com/bering-in-mind/) and *The Guardian* (http://www.guardian.co.uk/science/neurophilosophy). Several institutes in the United States also manage blogs that are often connected to a Facebook page or a Twitter account. Examples include *The neuroethics Blog* (http://www.theneuroethicsblog.com/) by the Center for Ethics at Emory University or *Mind the Gap* by the Center for Neuroscience and Society of the University of Pennsylvania (http://pennmindsthegap.wordpress.com/). Fewer examples can be found in Europe also because the search was limited to initiatives in English. The *neuromedia corner* represented an original attempt to provide sound information and foster the discussion on specific topics of neuroscience. It offered a lively website and a promising presence on Facebook and YouTube which, unfortunately, could not be explored and developed further due to the limited duration of the project. *The Practical Ethics Blog*

(http://blog.practicalethics.ox.ac.uk/) by the University of Oxford is the main example of blog at an institutional level, but some popular single-author blogs also exist.

The initiatives discussed in this chapter clearly represent a limited selection of major projects and key players at the European level. Many other relevant initiatives probably exist at a national level but are difficult to identify because of the language barrier.

The effects of activities of this kind emerge in a recent Eurobarometer Survey that suggests a "new era" in the relations between science and society ("Europeans and Biotechnology in 2010," 2010; Gaskell et al. 2011). According to this survey, European citizens show less criticism and more enthusiasm for new technologies even though they expect to see appropriate regulation for their use and ask for public involvement when social and ethical values are at risk. They also show general optimism for brain and cognitive enhancement research and even an approval of research on human enhancement, if kept under control. The survey investigated also the different influences that shape the relationship between science and ethics showing a complex pattern and a variability across countries which may hamper the definition of regulations at the European level and a true involvement of citizens in the process of policy-making. The Eurobarometer looked, in particular, at the influence of religion and education and identified different clusters of countries depending on the public's position on two crucial issues: whether science should prevail over science or vice versa and the "distributional fairness," i.e., whether the research may "benefit some people but put others at risk." These are crucial issues for neuroethics too and need to be addressed.

An excellent opportunity will be the "Year of the Brain" in 2014, a project promoted by the European Brain Council with "the ultimate goal to change and improve the way people think about the brain" (http://www.europeanbraincouncil.org/projects/eyob/). It includes the development of an interactive road show, school, and university program, a comprehensive website with features and articles for all forms of media, from TV to social networks. The project has already received enthusiastic support within Europe but has also generated interest outside the continent. Thus, the "Year of the Brain" in Europe 2014 has evolved into a three-year campaign called "The Age of the Brain", including the "Year of the Brain" in North America 2015 and the "Year of the Brain" in the Asia Pacific 2016. This will be a unique opportunity worldwide to foster awareness and engagement on brain science and the impact it may have on individuals' life and society. Hopefully, it will also be a starting point for more initiatives at the political and community level.

Cross-References

▶ Ethical Implications of Brain–Computer Interfacing
▶ Ethical Implications of Brain Stimulation

▶ Mind Reading, Lie Detection, and Privacy
▶ Neuroenhancement
▶ Neuroimaging Neuroethics: Introduction
▶ Neurolaw: Introduction
▶ Neuroscience, Neuroethics, and the Media

References

Bandelli, A., & Konijn, E. (2011). An experimental approach to strengthen the role of science centers in the governance of science. In J. C. Marstine (Ed.), *The Routledge companion to museum ethics* (pp. 164–173). New York: Routledge.

Europeans, science and technology. (2001). European Commission. http://ec.europa.eu/research/press/2001/pr0612en-report.pdf. Accessed 30 Dec 2012.

Europeans and Biotechnology in 2010. (2010). European Commission. http://ec.europa.eu/research/science-society/document_library/pdf_06/europeans-biotechnology-in-2010_en.pdf. Accessed 30 December 2012.

Gaskell, G., Allansdottir, A., Allum, N., Castro, P., Esmer, Y., Fischler, C., Jackson, J., Kronberger, N., Hampel, J., Mejlgaard, N., Quintanilha, A., Rammer, A., Revuelta, G., Stares, S., Torgersen, H., & Wager, W. (2011). The 2010 Eurobarometer on the life sciences. *Nature Biotechnology, 29*(2), 113–114.

Goldschmidt, R., & Renn, O. (2006). Meeting of minds – European Citizens' Deliberation on brain sciences. Final report of the external evaluation. Resource document. King Baudouin Foundation. http://www.kbs-frb.be/uploadedFiles/KBS-FRB/Files/Verslag/ECD_Finalreport_ExtEval_complete.pdf. Accessed 31 May 2013.

Illes, J., Blakemore, C., Hansson, M. G., Hensch, T. K., Leshner, A., Maestre, G., Magistretti, P., Quirion, R., & Strata, P. (2005). International perspectives on engaging the public in neuroethics. *Nature Review Neuroscience, 6*(12), 977–982.

Jensen, E., & Buckley, N. (2012). Why people attend science festivals: Interests, motivations and self-reported benefits of public engagement with research. *Public Understanding of Science* Published online on October 31, 2012. doi: 10.1177/0963662512458624.

Ramani, D., & Saviane, C. (2010). Neuroscience: experience of an interdisciplinary dialogue. *PCST 2010 Proceedings*. Retrieved from http://www.neuromedia.eu/UserFiles/file/RamaniSavianepaper.pdf.

Science and Society Action Plan. (2002). Resource document. European Commission. http://ec.europa.eu/research/science-society/pdf/ss_ap_en.pdf. Accessed 30 Dec 2012.

Traumatic Brain Injury and the Use of Documentary Narrative Media to Redress Social Stigma

95

Timothy Mark Krahn

Contents

Abstract

This chapter takes as its focus the topic of traumatic brain injury (TBI) from the perspective of "Public and Cultural Neuroethics," showing that there are deeply entrenched social and cultural barriers implicated in the trauma and injury commonly experienced by TBI survivors. The first section ("Traumatic Brain Injury (TBI) and the Situational Context for Survivors") details some basic facts about TBI and the situational context for this patient population. The second section ("Mistaken or Misattributed Beliefs About TBI") explains some of the prevalent societal misperceptions of persons living with TBIs, especially those linked to living with hidden disabilities as "the walking wounded." The third section ("Problems of Social Stigma") explains the problems of social stigma and what can be learned from on point mental illness research for making sense of societal challenges facing TBI survivors. The final section ("Educational Strategies and Using Narrative Media for Combating Stigma") concludes the chapter by examining educational strategies and the use of documentary narrative media for combating and redressing stigmas against this population.

T.M. Krahn
Novel Tech Ethics, Dalhousie University, Halifax, NS, Canada
e-mail: tim.krahn@dal.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_126,

Introduction

The human brain is a powerful organ and is commonly understood as a placeholder of privileged values, including those that are thought of as defining for our personal and human identity. Accordingly, philosopher Francisco Ortega and history of science professor Fernando Vidal have written on the rise of the “cerebral subject” in contemporary, industrialized societies, thereby referring to “an anthropological figure that embodies the belief that human beings are essentially reducible to their brains” (Ortega and Vidal 2007, p. 255). At least since the Enlightenment, a strongly influential strain in ethical theorizing has understood the human self to be a source or loci of unconditional value (Taylor 1989). If this is correct, and if as modern individuals we now are to understand human selves as essentially “cerebral subjects” (Vidal 2009), then it is easy to see how the human brain has come to assume its place as a source of *paramount* value in our culture(s). Ortega and Vidal’s recent book, *Neurocultures* (2011), maps the discourses, images, and practices that trace the social impact of the neurosciences and propel belief in the “cerebral subject” from public policy to the arts, from neuroscience to theology: they surmise that “[t]he new discipline of neuroethics is eminently symptomatic of such a situation” (Ortega and Vidal 2007, p. 255).

Indeed, neuroethics is celebrated as a new discipline, but as a new discipline – and influenced especially by philosophy – some scholars have seen fit to engage in self-reflexive investigations to determine its scope, purposes, and validity (Glannon 2007; Glannon 2011; Levy 2007; Levy 2012; Racine 2010; Roskies 2002). For those who see in the discipline a need for reflexivity (Brosnan 2011; Ahern 2011), considering the ethical implications of embracing or resisting the growing social impact of the neurosciences and creep of “neurocultures” means asking not only local questions about how this novelty might expand or contract what we already value. There is also a need to consider how all this focus on the brain (neuro-centrism) might create, or at least facilitate, certain barriers to overlooked or alternative courses of valuing and caring (see, e.g., Scully 2008) that do not sufficiently, but could perhaps better, command our present and future attention (cf. Racine and Forlini 2010; Outram and Racine 2011). An example of this reflexive approach that includes scoping what is on the horizon without “getting ahead of ourselves”[1] (so to speak) is Éric Racine’s *Pragmatic Neuroethics* (2010; see Racine 2008). Racine maintains that “the need for an interdisciplinary and collective response to ethical challenges in neuroscience *and* clinical care – neuroethics – has surfaced in the past years in response to important *social*, medical and scientific changes” [emphasis added] (Racine 2010, p. ix). One of the main

[1]In this regard, Racine’s work in neuroethics might be read as cutting off the path that slips into looking for answers to mental health in “neuroscience fiction” when relevant progress would seem more likely for the near future by paying greater attention to some of the methods, tools, and resources made available through the social sciences.

challenges for neuroethics, according to Racine – marking a new moral task that indicates a plurality of needs so complex and profound as to require a collective, integrated, interdisciplinary response (Racine 2008) – is the finding of the World Health Organization that the combined global "health burden" of neurological and mental health disorders now matches, if not surpasses, that of any cluster of health conditions (Racine 2010, p. ix; WHO 2006). If neuroethics is to play an important role in addressing the global health burden of neurological and mental health disorders, this importantly would have to be a matter not only for "Research Neuroethics" and "Clinical Neuroethics", but also one for "Public and Cultural Neuroethics" – defined by Racine as the "[e]thical challenges in the public understanding of neurological and psychiatric conditions: public engagement and the cultural representation of mental illness" (Racine 2010, p. 5).

This chapter takes as its focus the topic of traumatic brain injury (TBI) from the perspective of "Public and Cultural Neuroethics." If, as I am suggesting, there is some inflation – or misunderstanding (Glannon 2009:esp. ch.1) – going on with respect to the abovementioned tendency to see ourselves as defined essentially and singularly by our brains (Vidal 2009), what of those whose identities have been (at a minimum) shaken or, as the case may be, severely ruptured by a TBI? Plainly put, what is it to deal with a TBI in a culture with a matrix of social inscriptions fixing the view that "you are your brain" (Gazzaniga 2005, p. 31 qtd. in Ortega and Vidal 2007, p. 256), not simply that "you have a brain" (Vidal 2009)? Does the common experience of dramatic change in the survivor's life course and identity after a TBI (Widzinski 2009; Nochi 1998; Leith et al. 2004, p. 1204) prove correct the belief in the "cerebral subject"?

I want instead to show that there are not-so-new, well-worn, deeply entrenched social and cultural barriers implicated in the trauma and injury commonly experienced in the lives of TBI survivors. In this regard, documentary narrative media (e.g., video and film) is proving to be a promising tool in redressing these disabling, societal impediments.[2] As to the outline, the first section of this chapter ("Traumatic Brain Injury (TBI) and the Situational Context for Survivors") details some basic facts about TBI and the situational context for this patient population, focusing mainly on the USA. The second section ("Mistaken or Misattributed Beliefs About TBI") explains some of the prevalent societal misperceptions of persons living with TBIs, especially

[2]With the term "societal impediments" I would like to make room for understanding the barriers imposed by societies on TBI survivors as including more than just social factors. What is more, the disabling societal factors discussed in this chapter not only have a negative effect on the functioning of TBI survivors but also erect barriers against the possibility for "societies-in-general" from accessing the rich experiences and resources of this unnecessarily alienated population (cf. Scully 2008). Admittedly, this chapter focuses, in the main, on the losses experienced by TBI survivors, though more should be said on the relevant lost benefits to their respective societies (Sherry et al. 2010; cf. Swain and French 2000; The Disability Pride and Awareness Committee 2013).

those linked to living with hidden disabilities as "the walking wounded." The third section ("Problems of Social Stigma") explains the problems of social stigma and what can be learned from on point mental illness research for making sense of societal challenges facing TBI survivors. The final section ("Educational Strategies and Using Narrative Media for Combating Stigma") concludes the chapter by examining educational strategies and the use of documentary narrative media for combating stigmas against TBI survivors and for redressing problems arising from a "spoiled" social identity which marks this population.

Traumatic Brain Injury (TBI) and the Situational Context for Survivors

Depending on the site and extent of the insult (Levin 1993), TBI can involve effects on sensation, cognitive, emotional, motor, and behavioral functioning, (NIH 1998; Evans 2006; BIAC 2013) as well as personality changes sometimes complicated by lack of awareness (McAllister 2008, p. 4). Neurobehavioral sequelae vary in severity, length of time, and clinical manifestations (Riggio and Wong 2009, p. 163) leading to an "enormous range of outcomes" and varied needs on the part of the patient (Whyte 1998, pp. 23–24). Traumatic injuries to the brain are strongly associated with elevated risk of psychiatric conditions such as mood disorders (major depression, mania), anxiety disorders, psychosis, apathy, behavior or dyscontrol disorder, sleep disturbance, headache (Rao and Lyketsos 2000; Riggio and Wong 2009, p. 163), substance abuse (Taylor et al. 2003), and suicide (Teasdale and Engberg 2001; León-Carrión et al. 2012).

The WHO has recognized TBI as a universal public health concern (WHO 2006; McAllister 2008, p. 3; NINDS 2013), with a general incidence in developed countries commonly reported at 200 annually at risk per 100,000 population (Bruns and Hauser 2003). However, these rates are calculated solely on the basis of hospitalized cases and therefore do not include instances where injured persons either do not seek or have access to hospital care. Diagnosis is especially challenging, given that affected individuals may not manifest physical signs of injury. Studies in the past have shown mild TBI to be significantly underdiagnosed (NIHCDP 1999, p. 974). Even modern neuroimaging (e.g., MRI and CAT scans) or brain testing techniques (e.g., EEG) may fail to detect signs of brain injury (DVA 2013). McAllister surmises that "the actual incidence of injury is probably three to four fold larger than the [above] quoted numbers" (McAllister 2008, p. 3; cf. Thornhill et al. 2000).

Around the world, TBI is the foremost cause of death and neurologic disability in children and young adults (WHO 2006, p. 164). Thornhill and colleagues (Glasgow) have reported that approximately half of survivors admitted to hospital with mild or moderate TBI were classified as disabled a year afterwards and three quarters were so classified after a severe TBI (Thornhill et al. 2000). These authors conclude that "[t]he incidence of disability in young people and adults admitted with a head injury is higher than expected. This reflects the high rate

of sequelae previously unrecognised in the large number of patients admitted to hospital with an apparently mild head injury" (Thornhill et al. 2000, p. 1631).

Approximately 5.3 million persons in the USA (~2 % of the population) live with a TBI-related disability (CDC 2013), with an incidence rate estimated to be 150 annually at risk per 100,000 population (Kraus and Chu 2005). From 2002 to 2006 TBI-related emergency department visits in the USA increased by 14.4 % and hospitalizations by 19.5 % (Faul et al. 2010). Even though 75 % of recorded cases in the USA are mild forms of TBI, of 288,009 TBI survivors who were hospitalized in 2003, 124,626 (43 %) developed a long-term disability (Selassie et al. 2008).

Over the past four decades, in developed countries (including the USA), mortality rates have significantly declined with improvements in acute medical management (Arciniegas and McAllister 2008) and advances in trauma and neurosurgical services (McAllister 2008, p. 3; NIH 1998; Marshall 2000; Zink 2001). This has led to a concomitant increase in the population of persons living with disabilities associated with the neurobehavioral and neuropsychiatric disturbances resulting from TBIs (McAllister 2008; Arciniegas et al. 2000; McAllister 1992; Chesnutt et al. 1999, p. 9). Due to the complexity of the effects of insults on the brain, rate of recovery is often slow, with attendant high social, financial, psychological, and medical costs (McClure 2011, p. 85). As the 2006 WHO public health report *Neurological disorders: Public health challenges* maintains:

> The importance of rehabilitation is consistently underestimated, not least because of its cost. It is a regrettable truth that this part of the treatment lacks the drama of the primary treatment and is consequently more difficult to fund. It is nonetheless of great importance since TBI damages young lives for whom rehabilitation is as important for the regaining of function as primary treatment is for the saving of life. (WHO 2006:170)

Given the enormous range of outcomes possible with TBI, the goals of rehabilitation are very broad (Whyte 1998, pp. 23–24). Traditionally, in the USA, the focus has been "quite narrowly centred on acute medical restoration" (Leith et al. 2004, p. 1191; NIHCDP 1999). More recently, there has been a greater shift towards post-acute, community-based service delivery with the aim of integrating medical with psychosocial and environmental supports (Sloan et al. 2004; Willer and Corrigan 1994; Leith et al. 2004; Hibbard et al. 2002). However, as the need for services and supports becomes less and less like that which medical institutions are set up to deliver, funding for these unmet needs becomes ever more scarce (Whyte 1998, pp. 23–24). For instance, across the state healthcare systems in the USA, inpatient trauma and rehabilitation services tend to be reimbursed, while a large portion of outpatient services to assist community integration and facilitate well-being are either not adequately funded or not funded at all (Reid-Arndt et al. 2010, p. 142). Sadly, it is not uncommon that "efforts towards integrated care . . . end once a person with TBI leaves the medical arena" (Leith et al. 2004, p. 1192). The hardships for TBI survivors and their families are further compounded by the fact that recovery is oftentimes a long and incomplete process. To address these concerns

and meet the challenges of this growing public health concern, there is obviously a need to grow stakeholder as well as general public support for this population (Bérubé 1998, 2002, 2003). As Reid-Arndt and colleagues explain:

> Substantial efforts to educate others regarding the variety of possible outcomes following TBI are essential, as problems predicting the long term sequelae an individual may experience following a TBI complicates efforts to envisage needed programmatic services and supports. It is this complex nature of TBI, indeed, the difficulty comprehending the broad spectrum of consequences of TBI, which makes advocates' task difficult. (Reid-Arndt et al. 2010:138)

Mistaken or Misattributed Beliefs About TBI

In spite of the prevalence and gravity of brain injury, research indicates that nonexpert health professionals and the general public hold inaccurate or inadequate knowledge about TBI (Chapman and Hudson 2010, p. 797; Swift and Wilson 2001; Redpath and Linden 2004, p. 867), with historically very little progress in these matters (Hux et al. 2006, pp. 547–548; Guilmette and Paglia 2004, p. 188). For example, a US study by Gouvier and colleagues published in 1988, replicated by Willer et al. in 1993, with a follow-up again by Guilmette and Paglia in 2004, revealed that a high proportion of study participants believed that TBI survivors only suffer temporary cognitive deficits and that complete recovery from head injury is possible, contrary to expert opinion (Swift and Wilson 2001, p. 150; Chapman and Hudson 2010, p. 797; Hux et al. 2006, p. 550).

A prevailing reason that the changes in personality and new life challenges of the survivor are so difficult for others to understand and accept is that a large proportion of the effects of TBI, unlike physical injuries, remain invisible or not readily open to view (BIAC 2013). What is more, some of the effects of TBI – e.g., fatigue, poor concentration, memory loss, speech difficulties, anxiety, mood swings, and even intermittent aggression – may sometimes be seen in people without brain injuries, though not to the same degree or frequency (Guilmette and Paglia 2004; Matthews 2000). Even so, as with other nonvisible disabilities, "people may believe that there is nothing wrong with the injured persons" (McClure et al. 2006, p. 1029; see Matthews 2000; Miller and Sammons 1999). Expectations of TBI survivors are then based upon their usually intact, overt physical appearances of being normal (Swift and Wilson 2001; McClure et al. 2006). In this way, survivors become members of the "walking wounded." As a result, people who are not familiar with brain injury and lack awareness of survivors' psychiatric challenges form expectations for survivors' behavior as if they were neurotypical, oftentimes misattributing those symptoms that are the result of injuries ["actual cause" (Kelley 1967)] to other factors or causes (Leith et al. 2004, p. 1204; McClure et al. 2011; McClure 2011).

In the early stages, physical wounds or bandages on the head may indicate the experience of an insult to the brain, but TBI sequelae normally persist well past the point when physical markers of head injury are still readily apparent

(McClure et al. 2006, pp. 1029–1030). The literature reports that in cases where survivors lack visible markers of TBI, the resulting cognitive and behavioral sequelae are oftentimes misattributed to the everyday exigencies of life, life stage (Osborn 1998; Phillips 2000), or personality of the individual (Manchester 2006; McClure et al. 2006; McClure 2011, p. 86). Understandably, it is a small step from this to the added inference that the effects will not persist in the future (McClure et al. 2011, p. 395).

Research has also shown that not only the invisibility of the injury, but also the "dimension on which the injured person's behaviour is seen as abnormal," count as salient factors that distinctly affect attribution patterns (McClure 2011, p. 87). While family and expert professionals who know the survivor's history will usually make attributions based on how normal the TBI-related behavior presents as compared to the pre-injury state, other observers (absent this knowledge) will commonly make attributions simply according to compliance of the presenting behaviors with cultural norms (McClure and Abbott 2009; McClure et al. 2011, p. 393). Swift and Wilson have reported that even some hospital staff, family, friends, and the general public who experienced "inappropriate" behaviors from TBI survivors and who knew them to be brain injured, nonetheless failed to consider TBI as a possible source of explanation for these behaviors (Swift and Wilson 2001, p. 159). Swift and Wilson (2001) do not give specific examples here (at p. 159) of the behaviors judged to be inappropriate. However, later in their article they explain that:

> Although the health professional, unlike the lay person, is aware in many cases that the patient presenting with problems has a history of brain injury, many misconceptions attributed to health professionals who were not experts in the field of brain injury were similar to those held by the general public. These included inaccurate beliefs about: time span and extent of recovery; ability to return to work; behavioural symptoms being unrelated to the brain injury; the interpretation of physiogenic symptoms as psychological; the misinterpretation of motivation problems as laziness; and trivializing symptoms and their impact.[3] (Swift and Wilson 2001:160)

Hence, misattribution and/or discounting the injury as an explanation for behavioral and psychosocial sequelae leads TBI survivors to feel misunderstood, compounding their experience of adversity (McClure et al. 2006, p. 1029; McClure 2011, p. 90). Misattribution to transient causes such as life stage (e.g., adolescence) or personality may feed into expectations[4] for a recovery that is unrealistically swift and complete (Swift and Wilson 2001; McClure 2011, p. 87) – a possible "setup" for the survivor's perceived sense of failure. Such a scenario can work to compound survivor distress and frustration (McClure et al. 2006; McClure 2011) with the risk of further contributing to their existing deficits, thus impeding the recovery process (Guilmette and Paglia 2004, pp. 183–184). An additional misconception across the lay public (Gouvier et al. 1988), less experienced healthcare professionals (Ernst et al. 2009), as

[3] *Cf.* Gouvier et al. (1988) and Springer et al. (1997) for discussions of misperceptions of TBI survivors by their family members.

[4] See McClure et al. (2011) and Stone and Colella (1996) for a discussion of how people's expectations affect outcomes for persons with invisible disabilities.

well as educators (Farmer and Johnson-Gerard 1997) is that recovery is largely contingent upon the amount of effort exerted (to this purpose) by the survivor (McClure 2011, pp. 86–87). Hux and colleagues have reported a dramatic increase since 1993 in the tested public's misperception of the feasibility of complete recovery from severe TBI:

> Specifically, the public was largely incorrect in recognizing that complete recovery from severe brain injury is not possible, regardless of a person's desire for or commitment to self-improvement. Although 84.66 % of the 1993 respondents did not endorse this misconception, only 27.99 % of today's public recognized the fallacy of this notion. (Hux et al. 2006:550)

In general, then, many people (including healthcare professionals and service providers) either do not recognize, or actively minimize, the diversity of resultant cognitive difficulties and the broad impact of TBI-related injuries on the physical, psychological, emotional, behavioral, and social realms of the TBI survivor's life (Leith et al. 2004, p. 1204; see Swift and Wilson 2001; Guilmette and Paglia 2004; Hooper and Callahan 2001; McClure 2011, p. 90).

Problems of Social Stigma

Physically surviving a TBI is just the beginning of what is usually a lifelong process of adjustment and accommodation (Miller and BIRC 2009, p. 65). Researchers have consistently viewed the psychosocial challenges of TBI – chronic challenges related to the survivor's cognition, emotional functioning, and behavior in connection with interpersonal relationships, school, or work (NIH 1998) – as the most problematic for rehabilitation and critical to successful post-injury reentry into the community (Morton and Wehman 1995). As time proceeds post-injury, the aforementioned diversity of sequelae (including depression, aggression, anxiety, and communication disorders) contributes to the constriction of the survivor's social networks (Linden et al. 2005, p. 1011) in tandem with increasing intensity of time spent with family (Corrigan 2006, para.6; see Jacobs 1988; Kozloff 1987; Brzuzy and Speziale 1997). As Corrigan notes: "This constriction is not optimal for the individual or the family, and both end up feeling isolated" (Corrigan 2006, para.6; see Morton and Wehman 1995; Wallace et al. 1998).

In this regard, the societal impediments faced by TBI survivors and their families are in no way new:

> To borrow from the literature on mental illness, the attitudes society holds towards the mentally ill affects the development of social relationships (Thomsen 1992; Morton and Wehman 1995). Individuals perceived as being different are often treated with distrust, a lack of tolerance, and fear. (Linden et al. 2005:1011; see Guilmette and Paglia 2004; Willer et al. 1991)

A recent report (from the UK) by Linden and Boylan (2010) found all participants in their qualitative study typically used negative labels (e.g., perceptions of aggressiveness, dependency, and general unhappiness) when asked to describe TBI survivors

(see Linden and Crothers 2006; Linden and McClure 2012). In New Zealand, McLellan and colleagues tested for explicit and implicit negative attitudes (within the community) towards TBI survivors and if the terminology used ("head" vs. "brain") injury exacerbated negativity. Participants rated a hypothetical young adult subject described as having a "brain"/"head" injury more negatively than the same subject described as having a limb injury due to a car accident experienced during childhood. According to McLellan and colleagues, the negative attitudes observed were:

> openly endorsing less desirable characteristics to a young adult who experienced a brain injury as a child compared with a young adult who did not. In particular, they were regarded as less mature, intelligent, flexible, polite, and employable. In addition . . .using the term "brain injury" resulted in the individual being judged more negatively than when the term "head injury" was applied to the same injury event. (McLellan et al. 2010:708)

General discrimination against persons with disabilities is a widespread phenomenon well documented in the literature, which is arguably more intense and with less protections for persons with "more-mental-than-physical" disabilities (see Martinson 1998). In the mental health literature, research shows that the public hold significantly more negative attitudes and disapprove more strongly of persons with psychiatric disabilities than persons with related conditions of physical illness (Corrigan and Watson 2002a, p. 17; Corrigan et al. 2000; Socall and Holtgraves 1992; Weiner et al. 1988; van 't Veer et al. 2006). Corrigan and Watson point out that a prevalent reason for this is the widespread public belief that, unlike physical disabilities, persons with mental illnesses are in control and therefore responsible for causing their disabilities (Weiner et al. 1988; Corrigan et al. 2000; cf. Linden et al. 2007; Redpath et al. 2010). Research studies have shown that respondents "are less likely to pity persons with mental illness, instead reacting to psychiatric disability with anger and believing that help is not deserved" (Corrigan and Watson 2002a, p. 17; Socall and Holtgraves 1992; Weiner et al. 1988; Albrecht et al. 1982). Also, one Canadian study of knowledge about the sequelae of minor head injury and whiplash showed that lay persons may be significantly more sympathetic to complaints of physical symptoms resulting from an automobile accident in contradistinction to expressions of the same for cognitive difficulties (Aubrey et al. 1989). In this regard, Schomerus and colleagues have noted that attempts to impress upon the public that mental disorders are illnesses "like any other" (with biological correlates) may have been successful in promoting the acceptance of medical treatment for such, but has seemingly done little to promote greater social tolerance and positive acceptance of persons with neurological and psychiatric conditions (Schomerus et al. 2012, p. 450). Thus, TBI survivors, their families and support teams would appear to face many similar social challenges as those confronted by relevant populations living with mental illnesses (Redpath and Linden 2004, pp. 862–863). As Byrne notes: "Mental illness, despite centuries of learning and the 'Decade of the Brain', is still perceived as an indulgence, a sign of weakness" (Byrne 2000, p. 65).

Simpson and colleagues undertook a cross-cultural qualitative research project and found a common experience of problems with stigma and social isolation for TBI survivors and their families from all three backgrounds studied: Italian,

Lebanese, and Vietnamese (Simpson et al. 2000). Indeed, stigma has been identified as an issue "particularly associated with mental disorders and brain injury" (Redpath and Linden 2004, pp. 862–863; Phelan et al. 2011).

According to Goffman's theory, stigma is an attribute, behavior, or reputation that is socially discrediting insofar as it causes an individual to be classified by others according to an undesired, rejected stereotype in contradistinction from an accepted, normal one (Goffman 1963, p. 5). Stigmatizers make assumptions, attribute undesired qualities, and affix labels that assign "differentness" to those stigmatized, thus serving to distinguish the stigmatized from the "greater" society. But the differences so entrenched are for Goffman a matter of perspective, not reality: stigma arises because of the gap between virtual social identity and actual social identity (Byrne 2000, p. 65; Theriot 2013, p. 122). As we have seen with the case of TBI survivors, the resulting "tainted" (or "spoiled-in-advance") social identity negatively affects people's willingness to form social relationships with individuals that bear the marks of stigma and can intensify survivors' experience of social discrimination, isolation, and withdrawal (Theriot 2013, p. 122; cf. Corrigan et al. 2000). But for stigmatization to be effective, the stigmatized individual must internalize society's devaluation, owning the negative stereotype, thus recognizing her/his "spoiled" identity as a social reality (Jacoby et al. 2005; WHO 2006, p. 20; Burris 2008, p. 474). Goffman argues:

> The central feature of the stigmatized individual's situation in life is a question of what is often, if vaguely, called "acceptance". Those who have dealings with him fail to accord him the respect and regard which the uncontaminated aspects of his social identity have led him to anticipate receiving; he echoes this denial by finding that some of his own attributes warrant it. (Goffman 1963:8; cf. Burris 2008:474)

As such, "[s]elf-stigma results in a loss of self-esteem, diminished self-efficacy, and increased reticence about participation in social interactions" (Holmes and River 1998, p. 232). Not only is it hard for TBI survivors to face a society that misunderstands and devalues them; the net effect of stigma is also isolating, making for a seriously debilitating web of societal impediments that have historically prevented persons with serious psychiatric and neurological conditions from seeking and securing employment, assistance with living, adequate housing, and necessary healthcare (Baun 2009; Corrigan et al. 2000; Corrigan and Watson 2002a; Holmes and River 1998; Yang et al. 2008).

Educational Strategies and Using Narrative Media for Combating Stigma

The negative consequences of mistaken or misattributed beliefs, as well as generalized effects of stigma relevant to TBI, argue for various educational measures (Guilmette and Paglia 2004; McClure et al. 2011, p. 395) – including those targeting the general public (Mayville and Penn 1998, p. 250) – in order to foster an environment more conducive to social inclusion and support for TBI survivors (Scior et al. 2013; Walker and Scior 2013). As Peter Byrne, cochairman

of media projects for the UK Royal College of Psychiatrists' *Changing Minds* anti-stigma campaign, notes, "the starting point for all target groups and at every level is education" (Byrne 2000, p. 67; see WHO 2006, p. 21). Every educational intervention, as Byrne further explains:

> must convince its target group of the importance of stigma/discrimination, challenge stereotypes in ourselves and others, and pursue the ongoing task of unravelling the nature of prejudice. These three separate tasks are summarised in the *Changing Minds* slogan: "Stop, think, understand". (Byrne 2000:67)

In the literature, it has been argued that education can provide what is needed by the lay public to make more informed decisions about mental illness (Corrigan 1998, p. 217). As Corrigan and Watson (2002a) note, several studies have shown that participation in programs that strategically provide information on mental illness resulted in the lessening of negative stereotypes and improved attitudes about persons with the relevant conditions (Corrigan et al. 2001c; Morrison et al. 1980). But as Byrne and Corrigan and Penn contend, "closing the knowledge gap" on discrediting psychiatric stigma is only part of the answer (Byrne 2000, p. 68; Corrigan and Penn 1999).

The degree of familiarity with mental illness is a major impact factor for stigmatizing attitudes and behaviors (Corrigan et al. 2001b). Not simply greater knowledge but also *greater experience* with mental illness has been significantly correlated with less frequent (expressed) desire for social distance (Angermeyer et al. 2004; Vezzoli et al. 2001; cf. van 't Veer et al. 2006). Accordingly, researchers have consistently found personal contact to be most effective at reducing stigma associated with mental illness (Burris 2008; Corrigan et al. 2001a, c; Kolodziej and Johnson 1996; Lyons and Ziviani 1995; Penn et al. 1994, 1999; Phelan and Link 2004). This approach is now informing the large-scale social marketing efforts of the (US) *National Consortium on Stigma and Empowerment* (Corrigan 2011) and England's *Time to Change* campaign as associated with mental illness (Evans-Lacko et al. 2012, 2013).

The findings of research into stigma related to TBI are in line with what has been given here as evidence for combating stigma of mental illness. A recent study by Redpath and colleagues confirms that research participants who knew someone with a TBI "expressed less prejudice and were more likely to interact with the protagonist than those who didn't know someone with TBI" (Redpath et al. 2010, p. 808). Moreover, healthcare professionals who knew a TBI survivor were more likely to express more helping behaviors towards other TBI survivors (Redpath et al. 2010). McLellan and colleagues observed that "when people have more knowledge about or experience with brain injury, they are less likely to endorse negative stereotypes" (McLellan et al. 2010, p. 705),[5] and only those persons tested who were unfamiliar with TBI demonstrated a negative implicit bias. These TBI researchers concluded this study, saying:

[5]McLellan and colleagues' study also showed that test subjects unfamiliar with TBI and its effects had unconscious negative attitudes towards patients with TBI (McLellan et al. 2010; referenced by Phelan et al. 2011, p. 178). Burris has noted that diseased-based stigma may operate, in the main, through automatic emotional reactions and that familiarity may serve to reduce these reactions like the process of desensitization through exposure can reduce evolutionary-based phobias (Burris 2008, p. 475).

> Initiatives that increase the public's familiarity with brain injury and realistic sequelae may decrease negative evaluations. . . .Our findings suggests [sic] that negative evaluations of people with brain injury are not necessarily underpinned by an automatic or deep-seated bias, but rather are due to an openly held belief that brain injuries result in personality changes or deficits that render a person as less desirable. The explicit nature of the bias means it is reasonable to speculate that such negative attitudes may be more easily open to education, communication, and therefore amelioration. (McLellan et al. 2010:708; Rydell et al. 2007)

Consumer contact[6] is proving to be a most promising strategy in redressing stigma associated with psychiatric and neurological conditions (including TBI), especially when it is "targeted, local, credible, and continuous" (Corrigan 2011, p. 824). In the disabilities literature, Walker and Scior's review notes that "several studies have found a positive effect of direct contact on attitudes towards those with intellectual disabilities" (Walker and Scior 2013, p. 2201; McConkey et al. 1993; Rillotta and Nettelback 2007; Roper 1990). But while direct contact is very valuable, it may prove extremely hard to achieve on a large scale (Walker and Scior 2013, p. 2201) unless many more individuals come forward in public places (e.g., their workplaces) to disclose their experience of living with a stigmatized condition. Obviously this prospect would be risky for these individuals, given the prevalence and high costs (e.g. discrimination, loss of status, or loss of opportunities) associated with social stigma related to psychiatric and neurological conditions (Watson and Corrigan 2005, p. 291). This has led some anti-stigma activists and researchers to strategize the use of relevant film footage or video clips of contact in order to reach larger audiences (Dougall et al. 2012; Knifton et al. 2010; Quinn et al. 2011; Corrigan et al. 2007).

In this regard, there is building evidence that mere exposure can be an effective way to change negative and prejudicial attitudes (Walker and Scior 2013, p. 2201; Pettigrew 1998; Pettigrew and Tropp 2006; Zajonc 2001). The literature demonstrates specific successes with indirect contact interventions to improve attitudes to persons with intellectual disabilities or mental illnesses (Walker and Scior 2013; Carsrud et al. 1986; Hall and Minnes 1999; Smedema et al. 2012; Clement et al. 2012). One study by Corrigan and colleagues showed that a filmed version of contact led to greater mental illness stigma improvement (with positive changes on measures of pity, avoidance, and segregation) than an educational videotape without the additional indirect contact piece (Corrigan et al. 2007). Another study by Reinke, Corrigan, and colleagues showed relevant parity of improvement whether contact was presented in vivo or on videotape (Reinke et al. 2004; cf. Faigin and Stein 2010).

The use of film or video may prove additionally useful for reaching members of the lay public who might otherwise not be comfortable with, or even afraid and distrustful of (Linden and Crothers 2006), educational formats that include direct contact with persons affected by TBI (cf. Ritterfeld and Jin 2006). Arguably, films have the capacity to put viewers in contact with "a surrogate reality without having

[6]See Allport (1954), Pettigrew (1998), Pettigrew and Tropp (2000), and Couture and Penn (2003) for explanations of the intergroup contact theory hypothesis according to which direct interaction between members of the "in-group" with the "out-group" is thought to improve prejudicial attitudes (Walker and Scior 2013, p. 2201). See also Dovidio et al. (2003) and Pettigrew and Tropp (2008).

to take part in it" (Horsley 2009, p. 2), providing an opportunity to deal with stigma in a way that is importantly nonthreatening (cf. Watson and Corrigan 2005, p. 282; Macrae et al. 1994). As one interviewee reflecting on the *Scottish Mental Health Arts & Film Festival* put it:

> I like film anyway but I just think it's important to have that ability to reflect on issues or situations that actually happen in real life, in a safe sort of way. I suppose it makes it a bit safer because you're watching it on a film because you're not watching a real scenario, but it's not real and you can go away and think about it. (Dougall et al. 2012:131)

Across the USA today, narrative video media and documentary film are being used as tools of indirect (yet personal) contact to reduce stigma and foster a more positive social identity for TBI survivors. Part of this upswing is attributable to the fact that new technological advances in digital filming and computer-based editing have made film vignettes and documentaries increasingly affordable and easier to view and produce. Video capabilities through the internet have also made these materials widely accessible (Volandes 2007, p. 680). Accordingly, the (US) Center for Disease Control Injury Center and CDC Foundation have partnered to sponsor *Heads Up TBI Film Festival*, an online collection of TBI testimonials. TBI stakeholder participants are encouraged to create a video and upload it to *YouTube* in order to (i) share their "stories, experiences, successes, challenges, goals, memories and the hopes that motivate [them] each day" and (ii) "join a supportive community and share [their] ideas" (CDC Foundation TBI 2013). The CDC maintains that the goal of the film festival is to give participants a chance to "lend [their] voices so that TBI is no longer a 'silent epidemic'" (CDC Foundation TBI 2013). At the time of writing, 58 personal videos are listed with *Heads Up*, and a general search of *YouTube* for the terms "brain injury" and "story" yielded about 49,400 results.

Video narratives in this way are serving TBI survivors to connect and learn from one another and to promote their public visibility across a potentially wide spectrum of viewers. This phenomenon could be read as being "in step with" social movements across minority groups who have historically used storytelling as an important political tool in assisting to share experiences "that otherwise are not easily understood or appreciated" (Young 1997, p. 71). The importance of first-person perspective reporting (Ward and Meyer 1999, p. 137) and self-advocacy for taking greater control of public understanding of minority status social group identity is well known (Baun 2009; Cort 1987; Wilson 1999). Arguably these stories are working to change the face of TBI and, to borrow the words of longtime TBI advocate Constance Miller, they are fitting into a wider movement "to elevate the status of TBI from one of pity and shame to one of admiration and respect" (Miller and BIRC 2009). In this regard, these video narratives are what feminist bioethicist Hilde Lindemann Nelson has theorized as counterstories – namely, stories that provide a means through which to resist imposed misperceptions and misattributions that diminish agency, stories of self-definition that are "developed in response to the twin harms of deprivation of opportunity and infiltrated consciousness" (Lindemann Nelson 2001, p. 9).

If, as Byrne explains, "part of coping with stigma is fighting stigma" (Byrne 2000, p. 70), then counterstories provide an appropriate analytic for making sense of the growing movement within TBI communities to provide narrative repair (Lindemann Nelson 2001, p. 20) as a way to cope with not just damaged brains but damaged identities. This theme traces through the video documentary, *Brain Injury Dialogues*, which has been written and produced by TBI survivor Rick Franklin and documentary moviemaker Lyell Davies. The film works at bringing the world of brain-injured persons, "often 'invisible' on several levels, into the consciousness of the general public" as well as other TBI survivors, their families, and caregivers (Widzinski 2009). As the filmmakers visit with other TBI survivors, viewers see and hear "how survivors learn to deal with life after brain injury by means of personal, medical, and even political strategies," with a particular focus on the latter. In the "Filmmaker Statement" (tab) posted on the *Brain Injury Dialogues*-dedicated website,[7] Franklin explains that:

> one of the key messages in this video is the need for community support for survivors. . . .A second message . . .this documentary offers is how essential it is for all of us to better recognize the needs of brain injury survivors. For this to happen, we as brain injury survivors (with the help of disability rights, non-injured advocates, etc.) need to organize ourselves and both demand the accommodations we need, and reach out to help those also afflicted with this kind of injury. (Davies and Franklin 2010a)

Indeed, as featured TBI self-advocate and disabilities scholar Mark Sherry notes: "Peer support is critical to sorting out identity and gaining recovery" for survivors (Sherry et al. 2010), thus ameliorating some of their needs for connectedness and belonging (Leith et al. 2004, p. 1204). In this regard, the featured survivor support meetings include dialogue that parallels the late 1960s women's movement and its consciousness raising efforts which found that "the personal is political" (Hanisch 2006, p. 4). Indeed the *Brain Injury Dialogues* shows that when TBI survivors share their experiences of frustration, unhappiness, and anxiety, it seems that wider patterns of social stigma and a lack of societal concern and respect for their rights structure these personal, once seemingly private, stories and problems (cf. Young 1990, pp. 153–154). In this film, political dimensions emerge through social dialogue across survivors' stories which are critically explored together, with many personal problems revealed as public issues (Galvin 2005, p. 409). The film builds to a political conclusion that even though "no two brain injuries are alike" and that services and supports need to be personalized (Davies and Franklin 2010b, p. 14), solidarity and collective action are paramount, notwithstanding survivors' individual differences. Building a positive, cohesive, social identity for TBI survivors is also portrayed as politically important for gaining proper public recognition of survivors and regaining

[7]The website features a DVD transcript (in English), a Spanish translation of the transcript, special interview footage with TBI self-advocate and disabilities scholar Mark Sherry, TBI "Links of Note" (tab), reviews of the film, a link to the *Brain Injury Dialogues* Facebook page, and posted comments. See http://www.braininjurydialogues.org/.

public control of their lives (Sherry et al. 2010, p. 3; cf. Corrigan 1998, p. 215; Ward and Meyer 1999, p. 134). Mark Sherry's closing words of the film are unmistakably a battle cry to make the personal political in order to surmount forces of social stigma in the dominant culture:

> We, we in our hands, hold our fate. If we don't stick together, if we don't advocate, if we don't build something that's, that's much bigger than my little stay in hospital. . . if we don't say 'this was a really important thing but sometimes the system let me down, and for the next person, I've gotta make sure this doesn't happen'. Then we're letting down the next person in the same bed that we were. And that person really has far more in common with you than you care to realize. And that person is your brother and sister. Because they, they really are you. And, you know, what, what we, what we don't think about. . . what we, what we're trained in society to think about is this; I'm not like. . . I'm not like the person who's. . . down. I'm not like the person who's homeless. I'm not like the person who's psychotic. I'm not like the person who's "epileptic", or whatever. You know, I want to say the absolute opposite. I want to say to anybody who ever hears me. "Crazy?", I'm crazy. Epileptic? I'm "epileptic". Using a catheter? I'm using a catheter. Unable to breath for yourself? I'm unable to breath for yourself. Anything abject, anything disrespected, that's me. That's bed number two. That's bed number three. That's bed number four. That's us. That's our family. If we don't stick together. . . we, we live or die by that. (Davies and Franklin 2010b:15–16)

The *Brain Injury Dialogues* is thus an attempt to foster a new social story and social identity for TBI survivors. TBI survivors, on account of having (most commonly) invisible disabilities, are sometimes able to pass (Goffman 1963) for normal and therefore escape, at least initially, automatic stigmatization. There may be social incentives to conceal their disability to avoid stigmatization, but the damaging consequences include failing to have others recognize and accommodate their non-neurotypical needs (McClure 2011, p. 88; see Livneh et al. 2001; Matthews 2000), not to mention internalizing all of the pressures of maintaining social appearances to the contrary. The new forms of documentary narrative media just mentioned are providing opportunities and support to TBI survivors for managing their public identities and surmounting a state of "discreditable" stigma and self-stigma (Goffman 1963). In other words, the new forms of narrative media, though still not without social risk (cf. Corrigan and Watson 2002b), are providing advocates with an opportunity for "coming out" and publicly claiming TBI survivor status unashamedly as persons with otherwise invisible disabilities. As Mark Sherry puts it: "Brain injury pride. . . that's a completely different way to look at brain injury. To not focus [on] what's been lost, or to feel like less of a person because of it. But instead to take pride in who we've become after our injuries" (Davies and Franklin 2010b, p. 14; cf. The Disability Pride and Awareness Committee 2013).

Conclusion

As we have seen in this chapter, there are risks of a "spoiled" and/or "negative" identity (see Goffman 1963) as per societal barriers attached to the experience of stigma, social distancing, withdrawal, and isolation of TBI survivors. In a culture that puts such emphasis on the view that "you are your brain" rather than "you have

a brain" (Vidal 2009; Ortega and Vidal (2011); Ortega and Vidal 2007), it may seem that fixing the brain will fix the person. But as we have seen in this chapter, much of what makes for the trauma and injury suffered by TBI survivors – much of that which stands in the way of them learning how to repair and perhaps embrace their ruptured identities – is socially constructed in the form of stigma directed against this population. Swift and Wilson have aptly pointed out that "in the long term, brain injury may be conceptualized as a state which requires a careful re-negotiation and adaptation of the identity of the afflicted individual (Judd and Wilson 1999) rather than a condition which can be cured or controlled" (Swift and Wilson 2001, pp. 152–153). The fact that TBI often involves "hidden disabilities" makes matters of adaptation particularly precarious (cf. Pellerin et al. 2011). It is a hope that neuroethics will help to echo the call, already present in the mental illness and intellectual disabilities literature, to acknowledge that our brains and persons are socially situated (DesAutels 2010) and thus grow responsibility beyond the individual and family unit to communities and the wider public in order to provide the needed adjustments and accommodations for this population (Davies and Franklin 2010b; Sherry et al. 2010). Importantly, securing the needed adjustments and accommodations is an interactive and interpersonal process. Some of the new documentary narrative media platforms as directed by survivors represent promising endeavors to begin redressing some of their experiences of social and self-stigma. As such, they are providing emancipatory opportunities for claiming a new kind of social story and public identity for TBI survivors.

Acknowledgements Thanks to the Novel Tech Ethics research team members for helpful discussions and feedback in relation to this project. Funding for this project was provided by the Canadian Institutes of Health Research, NNF 80045, States of Mind: Emerging Issues in Neuroethics.

Cross-References

▶ Neuroethics and Identity
▶ Neuroethics Beyond Traditional Media
▶ Neuroethics of Neurodiversity
▶ Neuroscience, Neuroethics, and the Media
▶ Strengthening Self-Determination of Persons with Mental Illness
▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Ahern, M. K. (2011). Self-reflexivity in the formulation of autonomy: An appeal from feminist cultural studies. *American Journal of Bioethics Neuroscience, 2*(3), 52–54.

Albrecht, G. L., Walker, V. G., & Levy, J. A. (1982). Social distance from the stigmatized. A test of two theories. *Social Science & Medicine, 16*(14), 1319–1327.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge: Addison-Wesley.

Angermeyer, M. C., Matschinger, H., & Corrigan, P. W. (2004). Familiarity with mental illness and social distance from people with schizophrenia and major depression: Testing a model using data from a representative population survey. *Schizophrenia Research, 69*(2–3), 175–182.

Arciniegas, D. B., & McAllister, T. W. (2008). Neurobehavioral management of traumatic brain injury in the critical care setting. *Critical Care Clinics, 24*(4), 737–765.

Arciniegas, D. B., Topkoff, J., & Silver, J. M. (2000). Neuropsychiatric aspects of traumatic brain injury. *Current Treatment Options in Neurology, 2*(2), 169–186.

Aubrey, J. B., Dobbs, A. R., & Rule, B. G. (1989). Laypersons' knowledge about the sequelae of minor head injury and whiplash. *Journal of Neurology, Neurosurgery and Psychiatry, 52*(7), 842–846.

Baun, K. (2009). The role of the media in forming attitudes towards mental illness. *Moods Magazine* (www.moodsmag.com)*, Winter* (2009), pp. 27–29.

Bérubé, J. E. (1998). Brain injury advocacy. *Journal of Head Trauma Rehabilitation, 13*(5), 99–102.

Bérubé, J. E. (2002). Funding brain injury programs: The federal appropriations process. *Journal of Head Trauma Rehabilitation, 17*(1), 62–65.

Bérubé, J. E. (2003). A campaign for a "cure". *Journal of Head Trauma Rehabilitation, 18*(2), 204–206.

BIAC (2013). What is brain injury? Brain Injury Association of Canada. Retrieved 26 July 2011, from: http://biac-aclc.ca/en/

Brosnan, C. (2011). The sociology of neuroethics: Expectational discourses and the rise of a new discipline. *Sociology Compass, 5*(4), 287–297.

Bruns, J., & Hauser, W. A. (2003). The epidemiology of traumatic brain injury: A review. *Epilepsia, 44*, 2–10.

Brzuzy, S., & Speziale, B. A. (1997). Persons with traumatic brain injuries and their families: Living arrangements and well-being post injury. *Social Work in Health Care, 26*(1), 77–88.

Burris, S. (2008). Stigma, ethics and policy: A commentary on Bayer's "stigma and the ethics of public health: Not can we but should we?". *Social Science & Medicine, 67*(3), 473–475.

Byrne, P. (2000). Stigma of mental illness and ways of diminishing it. *Advances in Psychiatric Treatment, 6*(1), 65–72.

Carsrud, A. L., Ahlgren, R. D., & Dood, B. G. (1986). Evaluating the effects of a community awareness programme on attitudes towards sheltered work and living projects. *British Journal of Mental Subnormality, 32*(62, Pt.1), 37–41.

CDC (2013). Injury prevention & control: Traumatic brain injury. Centers for Disease Control and Prevention. from: http://www.cdc.gov/traumaticbraininjury/statistics.html

CDC Foundation TBI (2013). Heads up TBI film festival. Centers for Disease Control and Prevention. Retrieved 25 Jan 2013, from: http://www.youtube.com/user/CDCFoundationTBI/videos?view=0

Chapman, R. C. G., & Hudson, J. M. (2010). Beliefs about brain injury in Britain. *Brain Injury, 24*(6), 797–801.

Chesnutt, R. M., Carney, N. A., Oregon Health Sciences University, United States, & Agency for Health Care Policy and Research, (1999). *Rehabilitation for traumatic brain injury*. (no. 99-E 006 ed.) Rockville, MD: Agency for Health Care Policy and Research, U.S. Department of Health and Human Services, Public Health Service.

Clement, S., van Nieuwenhuizen, A., Kassam, A., Flach, C., Lazarus, A., de Castro, M., et al. (2012). Filmed v. Live social contact interventions to reduce stigma: Randomised controlled trial. *British Journal of Psychiatry, 201*(1), 57–64.

Corrigan, J. D. (2006). Consequences of traumatic brain injury for functioning in the community. NIH [National Institutes of Health] Consensus Development Conference on Rehabilitation of Persons with Traumatic Brain Injury (October 26–28, 1998, Bethesda, MD). Retrieved 28 July 2013, from: https://www.nichd.nih.gov/publications/pubs/TBI_1999/Pages/Abstracts.aspx#ConsequencesFunctioning

Corrigan, P. W. (1998). The impact of stigma on severe mental illness. *Cognitive and Behavioral Practice, 5*(2), 201–222.

Corrigan, P. W. (2011). Best practices: Strategic stigma change (SSC): Five principles for social marketing campaigns to reduce stigma. *Psychiatric Services, 62*(8), 824–826.

Corrigan, P. W., & Penn, D. L. (1999). Lessons from social psychology on discrediting psychiatric stigma. *American Psychologist, 54*(9), 765–776.

Corrigan, P. W., & Watson, A. C. (2002a). Understanding the impact of stigma on people with mental illness. *World Psychiatry, 1*(1), 16–20.

Corrigan, P. W., & Watson, A. C. (2002b). The paradox of self-stigma and mental illness. *Clinical Psychology: Science and Practice, 9*(1), 35–53.

Corrigan, P. W., River, L. P., Lundin, R. K., Wasowski, K. U., Campion, J., Mathisen, J., et al. (2000). Stigmatizing attributions about mental illness. *Journal of Community Psychology, 28*(1), 91–102.

Corrigan, P. W., Edwards, A. B., Green, A., Diwan, S. L., & Penn, D. L. (2001a). Prejudice, social distance, and familiarity with mental illness. *Schizophrenia Bulletin, 27*(2), 219–225.

Corrigan, P. W., Green, A., Lundin, R., Kubiak, M. A., & Penn, D. L. (2001b). Familiarity with and social distance from people who have serious mental illness. *Psychiatric Services, 52*(7), 953–958.

Corrigan, P. W., River, L. P., Lundin, R. K., Penn, D. L., Uphoff-Wasowski, K., Campion, J., et al. (2001c). Three strategies for changing attributions about severe mental illness. *Schizophrenia Bulletin, 27*(2), 187–195.

Corrigan, P. W., Larson, J., Sells, M., Niessen, N., & Watson, A. C. (2007). Will filmed presentations of education and contact diminish mental illness stigma? *Community Mental Health Journal, 43*(2), 171–181.

Cort, C. (1987). A long way to go: Minorities and the media. *Media & Values, 38*(Winter), online. Available at: http://www.medialit.org/media-values/media-values-articles-31-41

Couture, S. M., & Penn, D. L. (2003). Interpersonal contact and the stigma of mental illness: A review of the literature. *Journal of Mental Health, 12*, 291–305.

Davies, L. & Franklin, R. (2010a). *Brain injury dialogues*: Synopsis; filmmaker statement; DVD credits; DVD transcripts; links of note. Brain Injury Dialogues website. Retrieved 28 July 2013a, from: http://www.braininjurydialogues.org/statement.html

Davies, L. & Franklin, R. (2010b). DVD transcript: *Brain injury dialogues*. Brain injury dialogues website. Retrieved 28 July 2013b, from: http://www.braininjurydialogues.org/BID_Feature_English.pdf

DesAutels, P. (2010). Sex differences and neuroethics. *Philosophical Psychology, 23*(1), 95–111.

Dougall, R., Milne, R., Inglis, G., Onslow, H., Hesnan, J., & Knifton, L. (2012). Critical distance: How a mental health arts and film festival makes audiences think. *Arts & Health, 4*(2), 124–134.

Dovidio, J. F., Gaertner, S. L., & Kawakami, K. (2003). Intergroup contact: The past, present, and future. *Group Processes & Intergroup Relations, 6*(1), 5–21.

DVA (2013). Traumatic brain injury: A guide for patients. Department of Veterans Affairs, USA. Retrieved 9 June 2013, from: http://www.mentalhealth.va.gov/docs/tbi.pdf

Ernst, W. J., Trice, A. D., Gilbert, J. L., & Potts, H. (2009). Misconceptions about traumatic brain injury and recovery among nursing students. *Journal of Head Trauma Rehabilitation, 24*(3), 213–220.

Evans, R. W. (Ed.). (2006). *Neurology and trauma*. Oxford: Oxford University Press.

Evans-Lacko, S., London, J., Japhet, S., Rusch, N., Flach, C., Corker, E., et al. (2012). Mass social contact interventions and their effect on mental health related stigma and intended discrimination. *BMC Public Health, 12*(489), 1–8. doi:10.1186/1471-2458-12-489.

Evans-Lacko, S., Malcolm, E., West, K., Rose, D., London, J., Rusch, N., et al. (2013). Influence of *time to change*'s social marketing interventions on stigma in England 2009–2011. *British Journal of Psychiatry, 202*(Supplement), s77–s88. doi:10.1192/bjp.bp.113.126672.

Faigin, D. A., & Stein, C. H. (2010). The power of theater to promote individual recovery and social change. *Psychiatric Services, 61*(3), 306–308.

Farmer, J. E., & Johnson-Gerard, M. (1997). Misconceptions about traumatic brain injury among educators and rehabilitation staff: A comparative study. *Rehabilitation Psychology, 42*(4), 273–286.

Faul, M., Xu, L., Wald, M., & Coronado, V. (2010). *Traumatic brain injury in the United States: Emergency department visits, hospitalizations, and deaths*. Atlanta: Centers for Disease Control and Prevention.

Galvin, R. D. (2005). Researching the disabled identity: Contextualising the identity transformations which accompany the onset of impairment. *Sociology of Health & Illness, 27*(3), 393–413.

Gazzaniga, M. (2005). *The ethical brain*. New York: Dana Press.

Glannon, W. (2007). *Bioethics and the brain*. Oxford: Oxford University Press.

Glannon, W. (2009). Our brains are not us. *Bioethics, 23*(6), 321–329.

Glannon, W. (2011). *Brain, body, and mind: Neuroethics with a human face*. New York: Oxford University Press.

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs: Prentice-Hall.

Gouvier, W. D., Prestholdt, P. H., & Warner, M. S. (1988). A survey of common misconceptions about head injury and recovery. *Archives of Clinical Neuropsychology, 3*(4), 331–343.

Guilmette, T. J., & Paglia, M. F. (2004). The public's misconceptions about traumatic brain injury: A follow up survey. *Archives of Clinical Neuropsychology, 19*(2), 183–189.

Hall, H., & Minnes, P. (1999). Attitudes toward persons with down syndrome: The impact of television. *Journal of Developmental and Physical Disabilities, 11*(1), 61–76.

Hanisch, C. (2006). The personal is political - The women's liberation movement classic with a new explanatory introduction. www.carolhanisch.org. Retrieved 28 July 2013, from: http://www.carolhanisch.org/CHwritings/PIP.html

Hibbard, M. R., Cantor, J., Charatz, H., Rosenthal, R., Ashman, T., Gundersen, N., et al. (2002). Peer support in the community: Initial findings of a mentoring program for individuals with traumatic brain injury and their families. *Journal of Head Trauma Rehabilitation, 17*(2), 112–131.

Holmes, E. P., & River, L. P. (1998). Individual strategies for coping with the stigma of severe mental illness. *Cognitive and Behavioral Practice, 5*(2), 231–239.

Hooper, S. R., & Callahan, B. (2001). Traumatic brain injury: State of the state. *North Carolina Medical Journal, 62*(6), 336–339.

Horsley, J. (2009). *The secret life of movies: Schizophrenic and shamanic journey in the American cinema*. Jefferson: McFarland & Co.

Hux, K., Schram, C. D., & Goeken, T. (2006). Misconceptions about brain injury: A survey replication study. *Brain Injury, 20*(5), 547–553.

Jacobs, H. E. (1988). The Los Angeles head injury survey: Procedures and initial findings. *Archives of Physical Medicine and Rehabilitation, 69*(6), 425–431.

Jacoby, A., Snape, D., & Baker, G. A. (2005). Epilepsy and social identity: The stigma of a chronic neurological disorder. *The Lancet Neurology, 4*(3), 171–178.

Judd, D. P., & Wilson, S. L. (1999). Brain injury and identity – the role of counselling psychologists. *Counselling Psychology Review, 14*(3), 4–16.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (pp. 192–240). Lincoln: University of Nebraska Press.

Knifton, L., Quinn, N., Inglis, G., & Byrne, P. (2010). Ethical issues in a national mental health arts and film festival. *Journal of Ethics in Mental Health, 4*(2), 1–5.

Kolodziej, M. E., & Johnson, B. T. (1996). Interpersonal contact and acceptance of persons with psychiatric disorders: A research synthesis. *Journal of Consulting and Clinical Psychology, 64*(6), 1387–1396.

Kozloff, R. (1987). Networks of social support and the outcome from severe head injury. *Journal of Head Trauma Rehabilitation, 1987*(2), 3–14.

Kraus, J. F., & Chu, L. D. (2005). Epidemiology. In J. S. Silvers, T. W. McAllister, & S. C. Yudofsky (Eds.), *Neuropsychiatry of traumatic brain injury* (pp. 3–26). Washington, DC: American Psychiatric Press.

Leith, K. H., Phillips, L., & Sample, P. L. (2004). Exploring the service needs and experiences of persons with TBI and their families: The south Carolina experience. *Brain Injury, 18*(12), 1191–1208.

León-Carrión, J., De Serdio-Arias, M. L., Cabezas, F. M., Domínguez Rolda'n, J. M., Domínguez-Morales, R., Barroso, Y., Martín, J. M., et al. (2012). Recovery of cognitive function during comprehensive rehabilitation after severe traumatic brain injury. *Journal of Rehabilitation Medicine, 44*(6), 505–511.

Levin, H. S. (1993). Neurobehavioral sequelae of closed head injury. In P. R. Cooper (Ed.), *Head injury* (pp. 525–551). Baltimore: Williams & Wilkins.

Levy, N. (2007). *Neuroethics*. Cambridge: Cambridge University Press.

Levy, N. (2012). Neuroethics. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(2), 143–151.

Lindemann Nelson, H. (2001). *Damaged identities, narrative repair*. Ithaca/London: Cornell University Press.

Linden, M. A., & Boylan, A. M. (2010). 'To be accepted as normal': Public understanding and misconceptions concerning survivors of brain injury. *Brain Injury, 24*(4), 642–650.

Linden, M. A., & Crothers, I. R. (2006). Violent, caring, unpredictable: Public views on survivors of brain injury. *Archives of Clinical Neuropsychology, 21*(8), 763–770.

Linden, M. A., & McClure, J. (2012). The causal attributions of nursing students toward adolescent survivors of brain injury. *Nursing Research, 61*(1).

Linden, M. A., Rauch, R. J., & Crothers, I. R. (2005). Public attitudes towards survivors of brain injury. *Brain Injury, 19*(12), 1011–1017.

Linden, M. A., Hanna, D., & Redpath, S. J. (2007). The influence of aetiology and blame on prejudice towards survivors of brain injury. *Archives of Clinical Neuropsychology, 22*(5), 665–673.

Livneh, H., Martz, E., & Wilson, L. M. (2001). Denial and perceived visibility as predictors of adaptation to disability among college students. *Journal of Vocational Rehabilitation, 16*(3), 227–234.

Lyons, M., & Ziviani, J. (1995). Stereotypes, stigma, and mental illness: Learning from fieldwork experiences. *American Journal of Occupational Therapy, 49*(10), 1002–1008.

Macrae, C., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology, 67*(5), 808–817.

Manchester, D. (2006). Staff attributions for aggression and their relationship to treatment acceptability in brain injury rehabilitation. Thesis submitted for the degree of Doctor of Psychology, University of Surrey. Retrieved 28 July 2013, from: http://epubs.surrey.ac.uk/774/

Marshall, L. F. (2000). Head injury: Recent past, present, and future. *Neurosurgery, 47*(3), 546–561.

Martinson, N. (1998). Inequality between disabilities: The different treatment of mental versus physical disabilities in long-term disability benefit plans. *Baylor Law Review, 50*(2), 361–380.

Matthews, C. K. (2000). Invisible disability. In D. O. Braithwaite & T. L. Thompson (Eds.), *Handbook of communication and people with disabilities: Research and application* (pp. 405–421). Mahwah: Lawrence Erlbaum Associates.

Mayville, E., & Penn, D. L. (1998). Changing societal attitudes toward persons with severe mental illness. *Cognitive and Behavioral Practice, 5*(2), 241–253.

McAllister, T. W. (1992). Neuropsychiatric sequelae of head injuries. *Psychiatric Clinics of North America, 15*(2), 395–413.

McAllister, T. W. (2008). Neurobehavioral sequelae of traumatic brain injury: Evaluation and management. *World Psychiatry, 7*(1), 3–10.

McClure, J. (2011). The role of causal attributions in public misconceptions about brain injury. *Rehabilitation Psychology, 56*(2), 85–93.

McClure, J., & Abbott, J. (2009). How normative information shapes attributions for the actions of persons with traumatic brain injury. *Brain Impairment, 10*(2), 180–187.

McClure, J., Devlin, M. E., McDowall, J., & Wade, K. (2006). Visible markers of brain injury influence attributions for adolescents' behaviour. *Brain Injury, 20*(10), 1029–1035.

McClure, J., Patel, G. J., & Wade, K. (2011). Attributions and expectations for the behavior of persons with brain injury: The effect of visibility of injury. *Journal of Head Trauma Rehabilitation, 26*(5), 392–396.
McConkey, R., Walsh, P. N., & Conneally, S. (1993). Neighbours' reactions to community services: Contrasts before and after services open in their locality. *Mental Handicap Research, 6*, 131–141.
McLellan, T., Bishop, A., & McKinlay, A. (2010). Community attitudes toward individuals with traumatic brain injury. *Journal of the International Neuropsychological Society, 16*(4), 705–710.
Miller, C. & BIRC (2009). Brain injury resource center website. Head Injury Hotline. Retrieved 28 July 2013, from: http://www.headinjury.com/welcome.htm
Miller, N. B., & Sammons, C. C. (1999). People with nonvisible disabilities. In *Everybody's different: Understanding and changing our reactions to disabilities* (pp. 239–264). Baltimore: Paul H. Brookes Publishing Company.
Morrison, J. K., Cocozza, J. J., & Vanderwyst, D. (1980). An attempt to change the negative, stigmatizing image of mental patients through brief reeducation. *Psychological Reports, 47*(1), 334.
Morton, M. V., & Wehman, P. (1995). Psychosocial and emotional sequelae of individuals with traumatic brain injury: A literature review and recommendations. *Brain Injury, 9*(1), 81–92.
NIH (1998). Rehabilitation of persons with traumatic brain injury. NIH Consensus Statement (no.109) Online, October 26–28, 1998.16(1): 1–41. Retrieved 28 July 2013, from: http://consensus.nih.gov/1998/1998TraumaticBrainInjury109html.htm
NIHCDP. (1999). Rehabilitation of persons with traumatic brain injury. *NIH Consensus Development Panel on Rehabilitation of Persons with Traumatic Brain Injury: JAMA [Journal of the American Medical Association], 282*(10), 974–983.
NINDS (2013). Traumatic brain injury: Hope through research. Office of Communications and Public Liaison, National Institute of neurological disorders and stroke, National Institutes of Health. Retrieved 9 June 2013, from: http://www.ninds.nih.gov/disorders/tbi/detail_tbi.htm
Nochi, M. (1998). "Loss of self" in the narratives of people with traumatic brain injuries: A qualitative analysis. *Social Science & Medicine, 46*(7), 869–878.
Ortega, F., & Vidal, F. (2007). Mapping the cerebral subject in contemporary culture. *RECIIS - Electronic Journal of Communication, Information & Innovation in Health, 1*(2), 255–259.
Ortega, F., & Vidal, F. (Eds.). (2011). *Neurocultures: Glimpses into an expanding universe*. Frankfurt am Main: Peter Lang.
Osborn, C. L. (1998). *Over my head: A doctor's own story of head injury from the inside looking out*. Kansas City: Andrews McMeel Publishing.
Outram, S. M., & Racine, E. (2011). Developing public health approaches to cognitive enhancement: An analysis of current reports. *Public Health Ethics, 4*(1), 93–105.
Pellerin, C., Rochette, A., & Racine, E. (2011). Social participation of relatives post-stroke: The role of rehabilitation and related ethical issues. *Disability and Rehabilitation, 33*(13–14), 1055–1064.
Penn, D. L., Guynan, K., Daily, T., Spaulding, W. D., Garbin, C. P., & Sullivan, M. (1994). Dispelling the stigma of schizophrenia: What sort of information is best? *Schizophrenia Bulletin, 20*(3), 567–578.
Penn, D. L., Kommana, S., Mansfield, M., & Link, B. G. (1999). Dispelling the stigma of schizophrenia: II. The impact of information on dangerousness. *Schizophrenia Bulletin, 25*(3), 437–446.
Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65–85.
Pettigrew, T. F., & Tropp, L. R. (2000). Does intergroup contact reduce prejudice: Recent meta-analytic findings. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 93–114). Mahwah: Lawrence Erlbaum & Associates.
Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751–783.

Pettigrew, T.F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology, 38*(6), 922–934.
Phelan, J. C., & Link, B. G. (2004). Fear of people with mental illnesses: The role of personal and impersonal contact and exposure to threat or harm. *Journal of Health and Social Behavior, 45*(1), 68–80.
Phelan, S. M., Griffin, J. M., Hellerstedt, W. L., Sayer, N. A., Jensen, A. C., Burgess, D. J., et al. (2011). Perceived stigma, strain, and mental health among caregivers of veterans with traumatic brain injury. *Disability and Health Journal, 4*(3), 177–184.
Phillips, J. (Ed.). (2000). *The Everest within*. Auckland: Clarity Publishing.
Quinn, N., Shulman, A., Knifton, L., & Byrne, P. (2011). The impact of a national mental health arts and film festival on stigma and recovery. *Acta Psychiatrica Scandinavica, 123*(1), 71–81.
Racine, É. (2008). Interdisciplinary approaches for a pragmatic neuroethics. *The American Journal of Bioethics, 8*(1), 52–53.
Racine, É. (2010). *Pragmatic neuroethics: Improving treatment and understanding of the mind-brain*. Cambridge: MIT Press.
Racine, É., & Forlini, C. (2010). Cognitive enhancement, lifestyle choice or misuse of prescription drugs? *Neuroethics, 3*(1), 1–4.
Rao, V., & Lyketsos, C. (2000). Neuropsychiatric sequelae of traumatic brain injury. *Psychosomatics, 41*(2), 95–103.
Redpath, S. J., & Linden, M. A. (2004). Attitudes towards behavioural versus organic acquisition of brain injury. *Brain Injury, 18*(9), 861–869.
Redpath, S. J., Williams, W. H., Hanna, D., Linden, M. A., Yates, P., & Harris, A. (2010). Healthcare professionals' attitudes towards traumatic brain injury (TBI): The influence of profession, experience, aetiology and blame on prejudice towards survivors of brain injury. *Brain Injury, 24*(6), 802–811.
Reid-Arndt, S. A., Frank, R. G., & Hagglund, K. J. (2010). Brain injury and health policy: Twenty-five years of progress. *Journal of Head Trauma Rehabilitation, 25*(2), 137–144.
Reinke, R. R., Corrigan, P. W., Leonhard, C., Lundin, R. K., & Kubiak, M. A. (2004). Examining two aspects of contact on the stigma of mental illness. *Journal of Social and Clinical Psychology, 23*(3), 377–389.
Riggio, S., & Wong, M. (2009). Neurobehavioral sequelae of traumatic brain injury. *Mount Sinai Journal of Medicine, 76*(2), 163–172.
Rillotta, F., & Nettelback, T. (2007). Effects of an awareness program on attitudes of students without an intellectual disability towards persons with an intellectual disability. *Journal of Intellectual and Developmental Disability, 32*(1), 19–27.
Ritterfeld, U., & Jin, S. A. (2006). Addressing media stigma for people experiencing mental illness using an entertainment-education strategy. *Journal of Health Psychology, 11*(2), 247–267.
Roper, P. (1990). Changing perceptions through contact. *Disability, Handicap & Society, 5*(3), 243–255.
Roskies, A. (2002). Neuroethics for the new millennium. *Neuron, 35*(1), 21–23.
Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37*(5), 867–878.
Schomerus, G., Schwahn, C., Holzinger, A., Corrigan, P. W., Grabe, H. J., Carta, M. G., et al. (2012). Evolution of public attitudes about mental illness: A systematic review and meta-analysis. *Acta Psychiatrica Scandinavica, 125*(6), 440–452.
Scior, K., Addai-Davis, J., Kenyon, M., & Sheridan, J. C. (2013). Stigma, public awareness about intellectual disability and attitudes to inclusion among different ethnic groups. *Journal of Intellectual Disability Research, 57*(11), 1014–1026. doi: 10.1111/j.1365-2788.2012.01597.x
Scully, J. L. (2008). *Disability bioethics: Moral bodies, moral difference*. Lanham: Rowman & Littlefield Publishers.
Selassie, A. W., Zaloshnja, E., Langlois, J. A., Miller, T., Jones, P., & Steiner, C. (2008). Incidence of long-term disability following traumatic brain injury hospitalization, United States, 2003. *The Journal of Head Trauma Rehabilitation, 23*(2), 121–131.

Sherry, M., Davies, L., & Franklin, R. (2010). DVD transcripts from *the brain injury dialogues*: Mark Sherry with EBBI. Brain Injury Dialogues website. Retrieved 28 July 2013, from: http://www.braininjurydialogues.org/MS_Bonus_English.pdf

Simpson, G., Mohr, R., & Redman, A. (2000). Cultural variations in the understanding of traumatic brain injury and brain injury rehabilitation. *Brain Injury, 14*(2), 125–140.

Sloan, S., Winkler, D., & Callaway, L. (2004). Community integration following severe traumatic brain injury: Outcomes and best practice. *Brain Impairment, 5*(1), 12–29.

Smedema, S. M., Ebener, D., & Grist-Gordon, V. (2012). The impact of humorous media on attitudes toward persons with disabilities. *Disability and Rehabilitation, 34*(17), 1431–1437.

Socall, D. W., & Holtgraves, T. (1992). Attitudes toward the mentally ill: The effects of label and beliefs. *The Sociological Quarterly, 33*(3), 435–445.

Springer, J. A., Parmer, J. E., & Bouman, D. E. (1997). Common misconceptions about traumatic brain injury among family members of rehabilitation patients. *Journal of Head Trauma Rehabilitation, 12*(3), 41–50.

Stone, D. L., & Colella, A. (1996). A model of factors affecting the treatment of disabled individuals in organizations. *Academy of Management Review, 21*(2), 352–401.

Swain, J., & French, S. (2000). Towards an affirmation model of disability. *Disability & Society, 15*(4), 569–582.

Swift, T. L., & Wilson, S. L. (2001). Misconceptions about brain injury among the general public and non-expert health professionals: An exploratory study. *Brain Injury, 15*(2), 149–165.

Taylor, C. (1989). *Sources of the self: The making of the modern identity*. Cambridge: Harvard University Press.

Taylor, L. A., Kreutzer, J. S., Demm, S. R., & Meade, M. A. (2003). Traumatic brain injury and substance abuse: A review and analysis of the literature. *Neuropsychological Rehabilitation, 13*(1–2), 165–188.

Teasdale, T. W., & Engberg, A. W. (2001). Suicide after traumatic brain injury: A population study. *Journal of Neurology, Neurosurgery & Psychiatry, 71*(4), 436–440.

The Disability Pride and Awareness Committee (2013). Pride against prejudice: Disability awareness campaign begins November 5th. Inclusion Network. Retrieved 28 July 2013, from: http://www.inclusion.com/resdisabilitypride.html

Theriot, M. T. (2013). Using popular media to reduce new college students mental illness stigma. *Social Work in Mental Health, 11*(2), 118–140.

Thomsen, I. V. (1992). Late psychosocial outcome in severe traumatic brain injury. Preliminary results of a third follow-up study after 20 years. *Scandinavian Journal of Rehabilitation Medicine, 26*, 142–152.

Thornhill, S., Teasdale, G. M., Murray, G. D., McEwen, J., Roy, C. W., & Penny, K. I. (2000). Disability in young people and adults one year after head injury: Prospective cohort study. *BMJ [British Medical Journal], 320*(7250), 1631–1635.

van 't Veer, J. T., Kraan, H. F., Drosseart, S. H. C., & Modde, J. M. (2006). Determinants that shape public attitudes towards the mentally ill. *Social Psychiatry & Psychiatric Epidemiology, 41*(4), 310–317.

Vezzoli, R., Archiati, L., Buizza, C., Pasqualetti, P., Rossi, G., & Pioli, R. (2001). Attitude towards psychiatric patients: A pilot study in a northern Italian town. *European Psychiatry, 16*(8), 451–458.

Vidal, F. (2009). Brainhood, anthropological figure of modernity. *History of the Human Sciences, 22*(1), 5–36.

Volandes, A. (2007). Medical ethics on film: Towards a reconstruction of the teaching of healthcare professionals. *Journal of Medical Ethics, 33*(11), 678–680.

Walker, J., & Scior, K. (2013). Tackling stigma associated with intellectual disability among the general public: A study of two indirect contact interventions. *Research in Developmental Disabilities, 34*(7), 2200–2210.

Wallace, C. A., Bogner, J., Corrigan, J. D., Clinchot, D., Mysiw, W. J., & Fugate, L. P. (1998). Primary caregivers of persons with brain injury: Life change 1 year after injury. *Brain Injury, 12*(6), 483–493.

Ward, M. J., & Meyer, R. N. (1999). Self-determination for people with developmental disabilities and autism: Two self-advocates' perspectives. *Focus on Autism and Other Developmental Disabilities, 14*(3), 133–139. Fall.
Watson, A. C., & Corrigan, P. W. (2005). Challenging public stigma: A targeted approach. In P. W. Corrigan (Ed.), *On the stigma of mental illness: Practical strategies for research and social change* (pp. 281–295). Washington, DC: American Psychological Association.
Weiner, B., Perry, R. P., & Magnusson, J. (1988). An attributional analysis of reactions to stigmas. *Journal of Personality and Social Psychology, 55*(5), 738–748.
WHO (2006). Neurological disorders: Public health challenges. World Health Organization. Retrieved 10 Feb 2013, from: http://www.who.int/mental_health/neurology/neurological_disorders_report_web.pdf
Whyte, J. (1998). Rehabilitation of individuals with traumatic brain injury: Status of the art and science. In: Office of the Director National Institutes of Health (Ed.), *NIH Consensus Development Conference on Rehabilitation of Persons with Traumatic Brain Injury* (pp. 23–29). Bethesda, MD: National Institutes of Health. Available online from: https://www.nichd.nih.gov/publications/pubs/TBI_1999/Pages/Abstracts.aspx
Widzinski, L. (2009). Review of *brain injury dialogues*. Media Reviews Online. Retrieved 28 July 2013, from: http://libweb.lib.buffalo.edu/emro/emroDetail.asp?Number=3924
Willer, B., & Corrigan, J. D. (1994). Whatever it takes: A model for community-based services. *Brain Injury, 8*(7), 647–659.
Willer, B. S., Allen, K. M., Liss, M., & Zicht, M. S. (1991). Problems and coping strategies of individuals with traumatic brain injury and their spouses. *Archives of Physical Medicine and Rehabilitation, 72*(7), 460–464.
Willer, B., Johnson, W. E., Rempel, R. G., & Linn, R. (1993). A note concerning misconceptions of the general public about brain injury. *Archives of Clinical Neuropsychology, 8*(5), 461–465.
Wilson, N. L. (1999). Mental health and the media. *Journal of Humanistic Counseling, Education & Development, 38*(2), 68–69.
Yang, L. H., Cho, S. H., & Kleinman, A. (2008). Stigma of mental illness. In K. Heggenhougen (Ed.), *International encyclopedia of public health* (pp. 219–230). Oxford: Academic.
Young, I. M. (1990). *Justice and the politics of difference*. Princeton: Princeton University Press.
Young, I. M. (1997). *Intersecting voices: Dilemmas of gender, political philosophy, and policy*. Princeton: Princeton University Press.
Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*(6), 224–228.
Zink, B. J. (2001). Traumatic brain injury outcome: Concepts for emergency care. *Annals of Emergency Medicine, 37*(3), 318–332.

Section XX

Neurotheology

Neurotheology

96

Andrew Pinsent

Contents

Abstract

Neuroscience studies many of the conditions and concomitants for the exercise of those intellectual and moral powers usually considered most specific to the human person. The notion of the person is also central to what is called revealed theology. For this reason and others, prospects for neurotheology seem promising, despite the difficulties of demarcating the proper scope and methods of the field. An emerging theme is the special role of the non-dominant hemisphere in theology and religion. In addition, although strongly reductive interpretations that "explain away" theological matters are likely to attract ongoing interest, there are more expansive modes of interpretation. For example, the embodied cognition that is a central theme of modern neuroscience may have surprising parallels with a subtle, pre-Cartesian tradition of theological anthropology, offering new perspectives on human action and flourishing.

A. Pinsent
Ian Ramsey Centre for Science and Religion, University of Oxford, Oxford, UK
e-mail: andrew.pinsent@hmc.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_95,

Introduction

Neurotheology, which encompasses the study of matters at the intersection of neuroscience and theology (Newberg 2010), seems to offer intriguing prospects for research, but the challenges begin with demarcation of the field. The first definition of "theology" in the *Oxford English Dictionary* (Simpson 2013) begins, "The study or science which treats of God, His nature and attributes, and His relations with man and the universe; 'the science of things divine' . . .," the term "science" being intended in the classical and medieval sense of *scientia*, which pertains to knowledge that is demonstrated from first principles. By this definition, the subject matter of "theology" overlaps at least in part with that of philosophy. Aristotle remarks (*Metaph*. I, 2, 983a8–10), "God is thought to be among the causes of all things and to be a first principle" (Ross 1924), and "theology," "metaphysics," and "first philosophy," the study of first causes, were practically synonymous in his philosophy. Indeed, this close association is arguably also the reason why philosophical reflection on physics, the modern science that generally touches most closely on issues connected with "first causes," has long been regarded as offering a privileged opportunity for consideration of questions of "natural theology."

Although controversy remains about what, if anything, can genuinely be known about divine matters from the study of the natural world by natural means, there is no doubt that much of the core theological content of Christianity, to give an influential example, is considered within the perspective of faith to be unknowable even by the most refined exercise of human reason in the absence of a direct or indirect revelation from God. Such content is the subject matter of "revealed theology," the ordered study of what has purportedly been divinely revealed about God's relations to human beings and the universe. Among the objects in the world available for empirical study by science, it is embodied persons that are arguably most central to this revelation, since God purportedly communicates such matters either directly to persons or indirectly via other persons, for example, by testimony. Persons are also central to the metaphysical issues of theology. Indeed, this relationship is symbiotic, since the term "person" denotes a supreme principle and arguably the unique metaphysical principle to have emerged from revealed theology, originally to articulate the doctrines of the Trinity and Incarnation when the tools of classical philosophy proved inadequate (Spaemann 2006). Persons are also of central interest to the ultimate purposes of revealed theology according to its own terms, namely, a supernatural mode of human flourishing. This flourishing has been understood in much of the Western tradition, as articulated, for example, in the *Confessions* of St Augustine, as an interpersonal relationship with God culminating in divine friendship.

Since neuroscience provides new and powerful means for exploring what it means to be an embodied human person, there would appear to be considerable potential in the intersection of neuroscience with theological questions. In particular, neuroscience studies many of the conditions and concomitants for the exercise of those intellectual and moral powers usually considered most specific to the human person, shown not least in the existence of the discipline called "theology."

Challenges of Neurotheology

Despite the *prima facie* possibilities for a fruitful interaction of neuroscience and theology, considerable care is needed since neurotheology inherits a host of well-known intellectual hazards from other attempts to apply the findings of neuroscience to the humanities. Such risks include, for example, the temptation to mix the language of psychological powers such as "understanding" with those of the neural conditions and concomitants for the exercise of such powers. Additional potential for confusion, especially in theological matters, arises from the common tendency to transpose Cartesian language into the description of the brain-body relationship ("crypto-Cartesianism," cf. Bennett and Hacker 2003), along with the widespread but erroneous assumption that any view of the world that holds that human beings have immortal souls necessarily presupposes Cartesianism.

Another major challenge is the proper demarcation of a field devoted to the intersection of neuroscience and theology, not least because theology is often confused with religion and there is a well-known difficulty in defining "religion." For example, a focus of neurotheology is sometimes assumed to be the neuroscientific correlates of religious experience, but even a cursory glance at one of the influential works of theology, such as the *Summa theologiae* of Thomas Aquinas, shows that much of the discipline does not involve the study of purported experience, but reasoning about the complex consequences of such experiences in the world. Neurotheology might be better defined therefore as the discipline that addresses the neural conditions and concomitants involved in cognition, reasoning, and attitudes regarding specifically theological matters, which covers a vast range of possible topics. Yet it should be noted in passing that a broader sense of the field would also have to encompass topics that are usually studied independently of theology, especially in ethics, but in which theologians take a special interest or which take on subtly different meanings and priorities from a theological perspective. Some of the many examples include free will, responsibility, and consciousness. An even broader interpretation might encompass ethical questions arising from the treatment of the human person in the conduct of neuroscientific research.

Since even a "narrow" sense of neurotheology is extremely wide-ranging, the chapters of this section have been chosen to represent the key topics and styles of approach at the present time, ordered roughly in terms of increasing complexity of the theological subject material. Michael Trimble examines what can be said about the neuroscience of some kinds of basic religious experiences associated, for example, with sensed presences, acquisitions of information by certain atypical means, and distortions of time. Aku Visala reviews some of the development and multiple meanings of neurotheology, as well as providing a critical evaluation of the problems of translating the implications of neuroscientific work on basic religious experience into the complex conscious, volitional, emotional and experiential aspects of lived religious traditions. Adam Green examines some of the ways in which religious models or conceptual frameworks for understanding the world both shape and are shaped by cognitive processes. Iain McGilchrist draws from his extensive work on the asymmetric, complementary cognitive capacities of the two

hemispheres of the brain, examining implications about the inherent limitations of propositional descriptions and the value of images, metaphors, and narratives in acquiring, evaluating, and communicating religious knowledge that transcends what is already familiar. A final chapter examines how neuroscience can help interpret an account of ultimate human flourishing from a specifically theological perspective. In particular, this chapter examines recent work on the second person in experimental psychology and social neuroscience that has addressed a long-standing problem in moral theology, with implications for a new virtue ethics with a second-person perspective.

Emerging Themes in Neurotheology

Given the comparative novelty of the field, warnings about misinterpretation, and the range of complex issues covered, it might be unwise to express too many general conclusions about the findings of neurotheological research to date. Yet certain consistent themes seem to be emerging and are expressed in more detail in the chapters that follow.

The first of these themes coalesces around the recognition and importance of distinct modes of human cognition. The notion of distinct ways of grasping truth is not new of course, and can be discerned in the classical and medieval distinctions between *intellectus* (“understanding”) and *scientia*, noted previously, as well as distinct words in many languages for “to know,” such as *cognoscere* and *sapere* in Latin, and *kennen* and *wissen* in German. What neuroscience is adding, however, is insight into what happens when the exercise of certain cognitive powers is impeded or suspended, thereby helping to put their relative contributions on an empirical basis. McGilchrist, for example, draws attention to the multiple ways in which persons with right-hemisphere damage show empirically significant deficiencies in the recognition of wholes rather than parts (McGilchrist 2009). For instance, such persons may rely on a particular kind of part to identify a particular kind of whole (such the identification of a house by its chimney), and produce drawings that show a loss of overall coherence and integrity (Hécaen and de Ajuriaguerra 1952; Nikolaenko 2004). Right-hemisphere lesions can also result in an inability to recognize faces (prosopagnosia), and faces being described in curiously “flat” terms (Sergent and Villemure 1989). Such damage may also be correlated with loss of appropriate moral responses to situations, especially those that are associated with the perspective of a second person (Tranel 1994; Tranel et al. 2002; Damasio 1994).

The significance of these findings to neurotheology is that many of the matters that are experienced in religious practice and studied by theology – including liturgy, the use of metaphors and narratives in sacred texts, the prioritization of the face in many kinds of religious art, what is personal, living, implicit, and unexpected – seem to be most closely associated with the “non-dominant,” right hemisphere of the brain. By contrast, the left hemisphere prioritizes the abstraction of things from their contexts, the representations of things in terms of previously

established types and classes, modeling and consequent prediction, manipulation and utility. Of course, such generalizations are subject to all manner of standard caveats and both hemispheres in practice contribute to a wide spectrum of cognitive processes. Moreover, the priority of the right hemisphere in the certain matters of religious practice and theology does not deny the distinctive theological contributions of the left hemisphere, including abstract reasoning, representational modeling, and contemporary analytic theology. The point being underlined by neuroscience, however, is that many of the matters experienced in religion and studied by theology remain irreducible to the abstract reasoning most commonly associated with the left hemisphere, having particular need of what is suited to right-hemispheric cognition.

For those desiring to preserve and pass on their faith, a possible implication is that narrative, metaphor, art, and music will be at least as important as clear and distinct ideas and arguments, an important corrective, for example, to iconoclastic tendencies every few centuries in Christian history. Yet the direction of causation is unlikely to be one-way. Matters of theology and religious practice are not only cognized by the right-hemisphere cognition, but plausibly shape such cognition, and not only of religious matters. This influence is perhaps most clearly illustrated in art. There is a great loss of a sense of religious transcendence between Van Eyck, *Adoration of the Mystic Lamb* (1432) and Van Gogh, *Wheatfield with Crows* (1890), but the perception of nature is also remarkably altered between these paintings, paving the way for the more radical developments of art in the twentieth century. Just as the right hemisphere was long dismissed as a "minor," "silent," or even an inferior "regressing" organ (Henschen 1926; Oldfield 1966), but is now recognized as playing a crucial role in the perception of wholes and contexts, so too the matters of theology have often been dismissed in recent centuries as contributing nothing of factual relevance to the study of nature, but such dismissals miss the point. The communication of theological matters shapes the perceived organization of the world by individuals and societies, one possible reason why different religious traditions often correlate closely with profoundly diverse societies. Hence, the widespread dissolution or alteration of theological narratives and images in a society may have surprisingly profound effects on the practice and priorities of many other branches of knowledge, even those devoted to strictly scientific and empirical matters.

A second key theme that emerges from neurotheology takes the form of a bifurcation that parallels, in certain ways, the issue of the divided brain. On one hand, a considerable amount of scholarship as well as much popular writing on the interrelation of neuroscience and theology focuses on the conceptual reduction of neurological processes to simpler and more tractable models, drawn either from more basic, nonhuman organisms or electromechanical systems. One might loosely describe this approach in McGilchrist's terms as a left-brain perspective on the brain itself. Such modeling is extraordinarily useful for many phenomena, but runs the well-known risk of what Dennett has termed a "greedy reductionism" that oversimplifies (Dennett 1995), rendering opaque and perhaps also sparking hostility toward whatever does not fit neatly within the model. An example in the history of

neuroscience generally might be early enthusiasm for modeling the brain as essentially a digital computer (Dreyfus 1992). An example of a similar reductive eagerness in neurotheology may be popular interpretations of the so-called God-helmet experiments developed by Stanley Koren and Michael Persinger, who reported that many subjects had “mystical experiences and altered states” while wearing the apparatus (Persinger et al. 2010). Regardless of the ongoing debates about these experiments, there seems little doubt that much public interest was rooted in the intriguing possibility that religious experience, and ultimately the entire superstructure of theological reasoning, might ultimately be reduced to feelings that can, in principle, be induced artificially in the brain.

Strongly reductive interpretations, popularly interpreted as “explaining away” theology and religion, are likely to continue to attract popular interest. Nevertheless, the chapters of this section may also show that neuroscience and theology can interact in a way that one might broadly describe as expansive rather than reductive. To draw on an influential example noted previously, it can sometimes seem that there is an “unholy alliance” between body-soul dualism in recent centuries and body-brain dualism or “crypto-Cartesianism” today. Contemporary neuroscience can, however, also help to foster a more integral understanding of human being and action in ways that may have surprising parallels with pre-Cartesian theological anthropology. For instance, the extent of the unconscious contributions to human action underlined by contemporary neuroscience is challenging to dualism. By contrast, there are parallels with the more integral anthropologies of medieval moral theology, for which the soul is the “form” of the body, with a vast array of “passions” (*passiones*) and established dispositions contributing to action.

Conclusion and Future Direction

Since neuroscience provides new and powerful means for exploring what it means to be an embodied human person, and since persons are central to revealed theology, a *prima facie* optimism about neurotheology seems appropriate. Despite the difficulties of demarcating the proper scope and methods of the field, certain consistent themes also seem to be emerging. In particular, whether at the basic level of religious experience or the selection and use of appropriate metaphors for subtle theological issues, an emerging theme is the special role of the non-dominant hemisphere. In addition, although strongly reductive interpretations that “explain away” theological matters are likely to attract ongoing popular interest, there are more expansive modes of interpretation. For example, the embodied cognition that is a central theme of modern neuroscience may have surprising parallels with a subtle, pre-Cartesian tradition of theological anthropology. Provided the challenge can be met of translating between theological anthropology and neuroscience without loss or distortion of meaning, the future seems promising for an enriched understanding of what it means to be an embodied human person open to theological revelation, giving new perspectives on human action and personal flourishing.

Cross-References

- ► Consciousness and Agency
- ► Impact of Brain Interventions on Personal Identity
- ► Moral Cognition: Introduction
- ► Neuroethics and Identity
- ► Neuroscience, Free Will, and Responsibility: The Current State of Play

References

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Malden, MA: Oxford: Blackwell.

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Grosset/ Putnam.

Dennett, D. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

Dreyfus, H. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press.

Hécaen, H., & de Ajuriaguerra, J. (1952). *Méconnaissances et hallucinations corporelles: Intégration et désintégration de la somatognosie*. Paris: Masson et Cie.

Henschen, S. (1926). On the function of the right hemisphere of the brain in relation to the left in speech, music and calculation. *Brain, 49*(1), 110–123.

McGilchrist, I. (2009). *The master and his emissary: The divided brain and the making of the western world*. New Haven/London: Yale University Press.

Newberg, A. (2010). *Principles of neurotheology*. Farnham, Surrey, England/Burlington, VT: Ashgate.

Nikolaenko, N. (2004). Visual hemifield preferences: The cerebral hemispheres and the relationship to effective disorder. *Acta Neuropsychologica, 2*(4), 371–392.

Oldfield, R. (1966). Things, words and the brain. *The Quarterly Journal of Experimental Psychology, 18*(4), 340–353.

Persinger, M., Saroka, K., Koren, S., & St-Pierre, L. (2010). The electromagnetic induction of mystical and altered states within the laboratory. *Journal of Consciousness Exploration and Research, 1*(7), 808–830.

Ross, W. (1924). *Aristotle's metaphysics*. Oxford: Clarendon.

Sergent, J., & Villemure, J. (1989). Prosopagnosia in a right hemispherectomized patient. *Brain, 112*(4), 975–995.

Simpson, J. (ed.) (2013). Theology, n. *Oxford english dictionary online*. Oxford: Oxford University Press. http://www.oed.com/view/Entry/200388.

Spaemann, R. (2006). *Persons: The difference between "someone" and "something"*. Oxford: Oxford University Press.

Tranel, D. (1994). "Acquired Sociopathy": The development of sociopathic behavior following focal brain damage. In D. Fowles, P. Sutker, & S. H. Goodman (Eds.), *Progress in experimental personality and psychopathology research* (pp. 285–311). New York: Springer.

Tranel, D., Bechara, A., & Denburg, N. (2002). Asymmetric functional roles of right and left ventromedial prefrontal cortices in social conduct, decision-making, and emotional processing. *Cortex, 38*(4), 589–612.

The Contribution of Neurological Disorders to an Understanding of Religious Experiences

97

Michael Trimble

Contents

Abstract

Following a brief introduction to epilepsy and the historical link to religious experiences, several studies carried out at the Institute of Neurology on the potential neuroanatomical basis of such phenomena are reviewed. It is concluded that epilepsy, particularly temporal lobe epilepsy, is linked to religiosity, and that the nondominant hemisphere is particularly involved. The data are discussed in the light of the growing interest in such disciplines as neurotheology and neuroaesthetics.

M. Trimble
Institute of Neurology, London, UK
e-mail: mtrimble@ion.ucl.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_96,

Introduction

The last two decades or so has seen the development of a new neuroscience discipline, namely, neurotheology. The emphasis of the subject is the unraveling of the cerebral basis of religious experiences, an endeavor initiated by William James over 100 years ago, and now promoted by new methods of brain imaging and by other techniques such as external stimulation the human brain with noninvasive techniques. However, the spirit of neurotheology is by no means new, and physicians have written about the subject in one way or another since ancient times. The condition most linked with these contributions has been epilepsy, and studies with this condition will be discussed in this chapter. After a brief introduction to the subject, the history of the associations between epilepsy and religiosity will be briefly reviewed. This will be followed by an account of some of the scientific work in this area, including studies from the Institute of Neurology, London.

The Epilepsies

Epilepsy implies a tendency to have seizures which are repeated, and which are not solely the result of exogenous influences. The term is used above in the plural to emphasize that there are many different epilepsy syndromes, and to note that epilepsy reflects underlying syndromes of which seizures are the main, but not by any means the only clinical feature. The epilepsies are divided into two main groups. In one, there are those patients who have a focus or lesion in the brain, which can be assumed from the clinical description of the seizure or detected either with an EEG, or some form of brain imaging. Patients from this group are referred to as having lesion-related or focal epilepsies. The other main group, the generalized epilepsies, has a different presentation, primary variants having no obvious brain lesions, and sometimes an obvious genetic basis. Establishing a site of origin of the seizures in the generalized epilepsies is problematic, and they are often assumed to arise from subcortical generators affecting both sides of the brain simultaneously.

Many patients with focal epilepsy have what are referred to as partial seizures, these arising secondarily to a change in the structure or the function of one or several brain areas. In one from, which arises from the temporal lobes of the brain, the seizures can be rather specific in presentation, and in the past, the term "temporal lobe epilepsy" was used to refer to this epilepsy syndrome. In the current classifications, this is referred to as a form of lesion-related epilepsy.

Temporal lobe epilepsy has some defining characteristics, which include a typical seizure pattern, with specific signs and symptoms, often an identifiable change of structure and function in one or both temporal lobes of the brain, and a tendency to be difficult to treat.

Patients with temporal lobe epilepsy seem more likely to suffer from secondary disabilities, referred to as comorbidities, which include behavior disorders and classifiable psychiatric syndromes. Thus, patients with this form of disorder have an increased liability to develop depression and schizophrenia-like disorders, and, in

a subgroup, a special form of personality change has been described. This is named after two neurologists who more clearly formulated the features, Henri Gastaut from France and Norman Geschwind from America (hence the synonym the Gastaut-Geschwind syndrome), although clear descriptions of this personality anlage can be found in the English and continental medical literature of the 19.c. (Gastaut 1956; Trimble 1991, 2007b). For reasons that are not always clear, these behavioral syndromes found in some patients with epilepsy have occasioned much controversy. This is not discussed here, but the point to be made is that these clinical pictures represent forms of an organic brain syndrome, namely, a behavior disorder that evolves from the continuing presence of chronic brain dysfunction. It needs to be emphasized at this point that only a minority of patients with epilepsy develop symptoms and signs of the Gastaut-Geschwind syndrome, and the full syndrome is not necessarily found in all patients. Some have only some of the features.

The Gastaut-Geschwind Syndrome

The characteristics of the Gastaut-Geschwind syndrome include such symptoms as alterations in sexual behaviors, irritability, and viscosity, the latter being a tendency to slow, labored thinking, as if thoughts are emerging from treacle. This sometimes is revealed as circumstantiality. However, two fascinating features of the Gastaut-Geschwind syndrome that are relevant for this contribution are hypergraphia and hyper-religiosity.

The former was first clearly described by Geschwind and his colleagues, and refers to the tendency to write excessively and often compulsively (Waxmann and Geschwind 1974). They initially discussed seven patients, all of whom were hypergraphic and all of who had temporal lobe epilepsy. The writings of these patients were extensive, and characteristically meticulous, and in four patients the written themes had moral and religious overtones. In a further paper, they published three additional cases; one of the patients was reported to have undergone multiple religious conversions. The writing revealed a preoccupation with detail, and a "compulsive quality to much of the written output." Repetition of words, and often sentences, was seen.

Since their publications, there have been several other studies of this interesting neurological sign in patients with epilepsy (Roberts et al. 1982). Some have been clinical; others have used some form of rating scale to assess the extent of the writing, and the feelings of a need to write. Some patients display hypergraphia more in association with their seizures (post-ictally); in others, it waxes and wanes over long periods of time. It is probably an all or nothing phenomenon, rather than some graded trait, and, as such, is easily missed in small groups of patients. However, the results tend to confirm the assertions of the original investigators, namely, that hypergraphia is associated with temporal lobe epilepsy, especially from a site of seizure focus from the medial (i.e., limbic), structures. The question of the relationship of laterality to hypergraphia is at present unclear, although at least one report noted an overrepresentation of right-sided foci. While the relationship of the hypergraphia noted in epilepsy to the ability to produce creative written

text for the moment must remain speculative, the phenomenon is obviously the opposite to the effects noted with left hemisphere lesions, with the subsequent development of aphasia and agraphia.

It has been noted that the content of the writing from patients with epilepsy with hypergraphia often reflects religious or mystical themes. Of 15 published cases, nine had hyper-religiosity or comparable extensive metaphysical beliefs, and those who see these patients regularly in the clinic, and who ask about such phenomena, recognize the frequent association with the metaphysical.

Religiosity

The association between epilepsy and religion has been one that stretches back to antiquity. Oswei Temkin, in his brilliant account of the history of epilepsy, noted several ancient historical religious explanations for the illness (Temkin 1971). Either a god had sent it or a devil was assumed to have entered the patient, or the attacker was thought to have sinned against Selene, the goddess of the moon. Superstition has always veiled epilepsy, the old name for epilepsy, "the sacred disease," even lingering on today in some parts of the world. While at some times and in some places, the patient with epilepsy was considered to be unclean, touched by evil forces, and contagious, at others, he or she was considered divine, magic, and able to utter prophetic oaths.

Descriptions of patients with epilepsy whose religious feelings have been notably strong can be found in the writings of several authors. The case of the religious mystic Emmanuel Swedenborg is often quoted. Swedenborg was born in 1688 into a deeply religious family, he studied at the University of Uppsala, being mainly interested in mathematics. In 1774, he changed his career and life and became involved with the spiritual world and its understanding. He discoursed with angels, became chronically paranoid suffering from hallucinations and delusions, ultimately declaiming that Jesus Christ had made his second coming through him to found the Church of the New Jerusalem.

Swedenborg lived a solitary existence, barely washing, obviously eccentric, but writing profusely. At one time, when he was in London, he proclaimed he was the Messiah, and locked himself away for 2 days, emerging from his self-imposed incarceration with foam around his mouth. His case was written up by the famous English psychiatrist Henry Maudsley (1869), who referred to him as a learned and ingenious madman. Maudsley suggested that he probably had epilepsy, a theme taken up later by Foote-Smith and Smith (1996). They interpreted many of the descriptions in his writings as symptoms of temporal lobe epilepsy, including his states of ecstasy, which would be compatible with ecstatic auras.

In addition to Swedenborg's religiosity however, these authors postulated more extensive elements of the Gastaut-Geschwind syndrome within his personality, not only with a deepening of all of his emotions, but also with regards to his hypermoralism, his humorless sobriety, his mood lability, and his hypergraphia. They pointed out that his *Arcana Coelestia* was a work of more than two million

words. Swedenborg himself claimed that spirits dictated much of what he wrote, and what the spirits were dictating came from God.

Toward the end of the nineteenth century, there was a considerable expansion of both the clinical and experimental neurosciences, and these historical associations, between epilepsy and religion, attracted the attention of psychopathologists with an interest in these matters. An analogy was developed between the epileptic attack and the moments of inspiration of genius, and interestingly in this context, pride of place was given to a number of famous geniuses who were said, on account of their inspirations to have suffered epilepsy. Dostoievsky himself alikened his states of ecstasy with the experiences of Mohammad, and he included people with epilepsy in several of his novels, notably the character of Prince Myshkin in The Idiot.

The whole topic of the relationship between epilepsy and religion has been reviewed several times (Trimble 1991; Saver and Rabin 1997; Trimble and Freeman 2006). Saver and Rabin list the following religious people of whom it has been said, at one time or another, that they had epilepsy: St Paul, Mahomet, Margery Kempe, Joan of Arc, St Catherine of Genoa, St Teresa of Avila, St Catherine dei Ricci, Emanuel Swedenborg, Ann Lee, Joseph Smith, Dostoievsky, Heiroymous Jaegen, Vincent van Gough, and St Thérèse of Lisieux. The majority were suggested to have had partial seizures, complex in form and compatible with temporal lobe epilepsy.

One of the earliest of as new generation of studies was that of Dewhurst and Beard (1970), who described six patients that underwent religious conversions, all of which occurred post-ictally after a series of temporal lobe seizures, essentially forming part of a post-ictal psychosis. Here is a description of one of the conversion experiences from their series:

The patient’s first conversion experience occurred in 1955 at the end of a week in which he had been unusually depressed. In the middle of collecting fares (he was a bus conductor), he was suddenly overcome with a feeling of bliss. He felt he was literally in heaven. On admission to hospital, he said that he had seen God and that his wife and his family would soon join him in heaven; his mood was elated, his thought disjointed, and he readily admitted to hearing music and voices. He remained in this state of exultation, hearing divine and angelic voices for 2 days.

Personality Changes

The Gastaut-Geschwind syndrome defines a change of personality in people with epilepsy, mostly associated with chronic temporal lobe epilepsy, in which one of the cardinal features is hyper-religiosity. With such predispositions, the state of a psychosis is not reached, but patients may adopt a divine, religious lifestyle, often contrary to their previous habits, but if already of a religious inclination, they reveal an accentuation of their previous rituals and beliefs. Their mental state however may flower into a psychosis with a bout of seizures, when they will report the most intense religious experiences. Sometimes they kindle a belief, over time, that they have epilepsy for a specific purpose, namely, that they are special and have

been chosen to suffer with it, and it is God's gift to them. This can sometimes be very dangerous, since they then may harbor the belief that God will cure them if He wants to, and that their antiepileptic drugs are unnecessary, which they then stop taking. This leads to more seizures and a deterioration of the psychosis.

It seems likely that if the clinical descriptions and biographies of several of the people already noted above have any validity, they illustrate the development of such chronic hyper-religious states over time, and several of the personalities discussed would today perhaps would be recognized as exemplars of the Gastaut-Geschwind syndrome.

Studies Carried Out at the Institute of Neurology

In recent years, The Raymond Way Research Group at the Institute of Neurology, Queens Square, London, has conducted several studies specifically examining aspects of religiosity and the often associated hypergraphia in patients with epilepsy. In one of the first, six cases of patients with temporal lobe epilepsy and hypergraphia were examined, and data collected and combined with those from 15 other published cases (Roberts et al. 1982). The patients wrote long, often detailed descriptions of their lives, and several were producing religious texts, such as the Bible or the Koran. Poetry writing was often a feature. There was often a meticulous attention to detail, and an associated mood disorder, mainly of euphoria, the latter was often linked to the hypergraphia. One patient, a journalist, was hypergraphic in the course of his work, but post-ictally, associated with a change of mood, he started writing in an entirely different style, and the content became philosophically bound, in complete contrast to his every day journalism. In the series, hyper-religiosity or comparable metaphysical beliefs were reported in nine cases, and seven had episodes of déjà vu. No patient had a left-sided focus for the epilepsy, and a statistically significant excess of right temporal abnormalities was noted. Déjà vu auras are usually driven by a right temporal lobe focus.

Here is a description of the full syndrome, described by Bear (1986):

A 56-year-old woman experienced a foul odor, followed by focal movements of the left face and arm, which on occasion generalized to a tonic-clonic convulsion. The surface EEG and recordings from indwelling electrodes implicated a right temporal lobe spike focus localized to the amygdala. The patient had composed thousands of pages of handwriting, distinguished by somber personal reflections, religious exegeses, and angry diatribes against former physicians, police, and politicians. Her writing was especially noteworthy, because she suffered painful, deforming rheumatoid arthritis, which required the use of finger and wrist supports. In her diaries, she commented "my hands are so sore but I have to just write." In addition to the unmistakable religious fervor of the writing, she traveled with satchels of audio-cassettes containing her own sermons on biblical themes.

In a more recent study at the Institute of Neurology, carried out with Anthony Freeman, the phenomenology of these states was investigated, especially the religiosity, in more detail, and has attempted a further understanding of the underlying brain associations (Trimble and Freeman 2006).

One problem is that measurement of such phenomena as religiosity and hypergraphia is difficult, with few available validated rating scales, and mainly a tradition of clinical observation to guide hypotheses. One scale used is that developed by Bear and Fedio, which was devised at the US National Institute of Health, precisely to fill in this gap. It consists of questions to assess 18 personality traits, many linked in the literature with the Gastaut-Geschwind syndrome (Bear 1986). One of the subscales relates to religiosity, another to hypergraphia. The scale comes in two versions, one for the patient to fill out, and the other for a carer or relative or person who knows the patient well.

In the original study of Bear and Fedio, the scoring on the religiosity subscale was greater in patients with temporal lobe epilepsy, even when compared to patients with mixed psychiatric disorders (and no epilepsy). However, there has been a controversy as to the specificity of the findings for the temporal lobe epilepsy, in part because of difficulties of studying patients with other forms of epilepsy, and excluding a temporal lobe focus. In the study to be described, the Bear-Fedio Scale was used, in addition to other scales that assess aspects of individual religious experiences and behavior. One scale was used, referred to here as INSPIRIT, and another, namely, the Hood Mysticism Scale. The INSPIRIT is a questionnaire that asks about spiritual or religious beliefs and experiences, including time spent on various religious practices, and how close people have felt to powerful forces of one kind or another. There are also direct questions about belief in God, and experiences that may have reinforced such beliefs. In our studies, this scale was modified to allow for a wider range of religious experiences to be documented than in the original and better descriptions of the nature of them. The Hood scale, which was based on Hood's readings of James's varieties of religious experiences, taps into the quality of them. There are two major factors evaluated by this scale, namely, general mystical experience and religious interpretation (Hood 1975).

Three experimental groups were defined. The first consisted of 28 people with temporal lobe epilepsy and a prominent devotion to religion – identified clinically. The second consisted of 22 people with temporal lobe epilepsy who had no religious affiliations, while in the third group were 30 regular church-goers, without known epilepsy. The purpose of the study was to examine in more detail the psychological profile of those patients with epilepsy and religiosity, and, by comparing them with the other epilepsy sample, to examine the underlying epilepsy variables that may be related to the religious experiences. By examining a group of non-epileptic worshipers, we hoped to capture phenomenological differences between them and our epilepsy sample.

The results of this study were very revealing. First of all, the original findings of Geschwind and his school were reconfirmed and the temporal lobe religious group not only as expected, endorsed the religiosity subscale of the Bear-Fedio scale, but they also revealed other elements of the Gastaut-Geschwind syndrome. Notably, the religious group also scored highly on the subscales of emotionality, philosophical interests, anger and sadness, dependence, and hypergraphia. They were also rated by a significant other to have more paranoia and mental viscosity than the non-religious sample. Thus, the profile that emerged in the religious patients with epilepsy was true to the original clinical descriptions of the Gastaut-Geschwind

syndrome, and emphasized hypergraphia, philosophical interests, and emotionality linked with the hyper-religiosity.

When the two groups of patients with temporal lobe epilepsy were compared, those with the religiosity were noted to more often have had a history of episodes of post-ictal psychosis, and more electroencephalogram changes that were bilateral compared to those without the religiosity.

The ordinary church-goers were different in their backgrounds from the epilepsy patients (more females, of an older age who were more likely to be married, and with a better educational level and occupational status), but they also differed in their religious behaviors. The patients with epilepsy and religiosity were more likely to belong to a religion not regarded as mainstream (i.e., Church of England, or Catholic); Seventh Day Adventism being popular. The question : "How often have you felt as though you were very close to a powerful spiritual force that seemed to lift you outside?," was endorsed more by those with epilepsy and religiosity than the non-epilepsy churchgoers, but interestingly, there was no significant difference in response to the question that asked if they had experienced any religious experience. Neither was there any difference in questions which asked about the frequency or duration of the religious experiences that they had.

The following statement from the INSPIRIT were endorsed more by the religious when compared to the non-religious epilepsy group:

How often have you felt as though you were very close to a powerful spiritual force that seemed to lift you outside yourself?

Items also endorsed more by the religious epilepsy sample were:

"an awareness of an evil presence";
"an experience of a miraculous event";
"a sensory or quasi-sensory experience of a great spiritual figure" or "an experience of a great spiritual figure" (such as Jesus, Mary, Elijah, Buddha or Allah);
"an experience with near-death or life after death";
"an experience of being punished in same way by God";
and "an overwhelming experience of fear".

On the Hood Mysticism Scale, the main features that identified the religious epilepsy sample were the "noetic," "temporal/spatial," and "ineffable" qualities of their experiences. Noetic here refers to the experience itself as a valid source of knowledge, with an emphasis on non-rational, intuitive, insightful experiences that none-the-less are recognized as not being purely subjective. Ineffability means the inability to express the experience in ordinary language, while the temporal/spatial dimension reflects on the experience as one in which time and space are modified.

The patients with bilateral temporal lobe electroencephalographic changes in particular scored higher on:

"the sense of presence of evil";
"an experience of a great spiritual figure";
"an experience of angels";
"a sensory or quasi-sensory experience of angels";
"a near death experience";
"an overwhelming fear";

"complete joy and ecstacy";
"being punished by God";
and "loss of self control".

These results lead to some relevant conclusions. Firstly, temporal lobe epilepsy can be associated with an identifiable constellation of behavioral dispositions, which have been identified for years, but are now better documented with rating scale assessments. This in its full form is referred to as the Gastaut-Geschwind syndrome. Secondly, patients with epilepsy and hyper-religiosity often have experienced post-ictal psychotic states, underscoring the potential links between their psychological profile and that of an epilepsy related psychosis. It is of interest in this context that states of hyper-religiosity are often seen in the context of such a psychosis (Kannemoto 2002).

With regard to the actual religious experiences, the epilepsy religious sample reported more experiences of the feeling of the presence of some external being, either of evil or of great spirituality, associated with feelings of death or dying and intense fear. The experiences are ineffable and noetic, not simply an awareness. There is identification, of and with that essence, and ecstasy and miracle are features of the descriptions.

Some caution must be given to the interpretation of the results of the above study, especially comparing church-goers without epilepsy with the religious patients with temporal lobe epilepsy, in part because of the demographic differences. However, from the neurological point of view, it was the patients with temporal lobe epilepsy and bilateral disturbances of function (as revealed through the electroencephalogram) in their temporal lobes who reported the most intense religious experiences, especially endorsing items to do with the inner subjective nature of their experiences, their ineffability, and the spatial/temporal component.

Other groups have also studied the links between epilepsy and religiosity. Csernansky et al. (1990) looked for correlations between psychopathological and neurological variables in a small sample of patients with epilepsy. They did not find a statistical difference with regard to laterality, but favored the left and bilateral abnormalities as opposed to the right hemisphere in association to the religiosity. Some others, also using rating scales, have not found such positive associations. These results have been reviewed elsewhere (Trimble 1991, 2007), but they do raise caution in interpreting these results.

MRI Studies

Of several hypotheses that have now evolved regarding the cerebral substrates of religious experiences, a role of the nondominant hemisphere and of some limbic lobe structures seem to have emerged as both testable and valid within the context of the growing understanding of brain structure and function, especially with regard to emotion. The above-reviewed literature seems biased toward a predominant role of the nondominant hemisphere, and studies in biological psychiatry have implicated this hemisphere as the one more associated with some affective disorders, namely, bipolar disorders and hypomania. Further, epilepsy has been shown to be

an interesting and reliable model of a neurological disorder associated with religiosity. It was with this background that the volumes of the hippocampus and the amygdala were measured using MRI in patients with epilepsy whose interictal behavior was rated using the Bear-Fedio Inventory (Wuerfel et al. 2004).

Thirty-three patients with refractory epilepsy were examined, and the volumes of the limbic structures were compared between those who scored high or low on three subscales of the Inventory, namely, "religiosity," "writing," and "sexuality." Patients who met the criteria for hyper-religiosity had significantly lower mean right hippocampal volumes than those not reported as religious. While in itself this finding was interesting, confirmation of the association emerged when the actual scores on religiosity were correlated with hippocampal size. Thus, there was a significant negative association between religiosity scores and hippocampal size, on the right side only. Further, this correlation was found for both patient self-ratings and on the independently obtained carer ratings – the smaller the right hippocampus, the greater the expressed religiosity. These data suggested that the right limbic structures may be of central importance to the development of these personality attributes, in keeping with the above developed hypothesis.

Other Neurological Disorders

Although it is the case that neurological conditions other than epilepsy are associated with states of religiosity, in reality they are few, and little discussed in the neurological literature. The best overall review is that of Saver and Rabin (1997). They point out that most religious experiences parallel similar non-religious ones, such as joy, love, fear, and awe, and that the emotions associated with the religious experience are ordinary emotions, differing little, if at all, in their emotional tone, but substantially differing in being directed toward a religious object.

They point out that the expressions of any particular religious language will depend upon the same mechanisms within the brain that relate to the expression of non-religious language, and conclude that the cerebral circuits underlying religious affect and cognition are not only widely distributed in the human brain but are a part of the cerebral apparatus available to everyone, the ecstatically religious and the non-religious alike. Thus, they opine that what is peculiarly distinctive to the religious experience is not so much to be found in the realm of affect, language, or cognition but in perception, namely, the direct sensory awareness of God or some other being. They, as William James, did not postulate an identifiable separate organ of religious perception, and in their paper, they attempt to understand the neurological substrates of such experiences as the direct awareness of a sacred or divine presence.

The above cited work on epilepsy has emphasized the links between temporal lobe-limbic structures and such experiences, and the personal descriptions of patients with epilepsy who are religious have revealed the intensity of these sensations. With regard to other neurological conditions, frontotemporal dementia (FTD) is almost the only other neurological condition discussed in the context of religiosity.

Frontotemporal Dementia

Dementia, which refers to a number of neuropathologies in which there is an acquired loss of intellectual abilities, has many causes, from the static, such as a head injury or injury to the brain from loss of oxygen, to the progressive. The latter dementias, sometimes referred to as parenchymatous dementias, usually arise in mid to late life, and present with a slowly progressive loss of skills and intellectual capacity, inevitably leading to dependency and death. The best known form is Alzheimer's disease, and a second common variety is dementia secondary to cerebrovascular disease. A third form, namely, frontal dementia, has been described for many years, and was referred to at one time as Pick's disease.

It has now become appreciated that frontal types of dementia are much commoner than previously thought, and the clinical and pathological pictures extend far beyond that condition described by the neuropathologist Pick. Thus, as a group, the FTDs tend to affect the frontal and temporal lobes of the brain and the presenting features are usually behavioral rather than cognitive in the first instance. The clinical picture will vary, depending on which parts of the brain are first affected by the pathology and on the rate of progression of the disease. In some patients, unilateral frontal or temporal lobe pathology occurs, and the temporal variant has been well described by Bruce Miller and his group from UCLA (Edwards-Lee et al. 1997). Hyper-religiosity is a feature:

Patient RTLV: A 59-year-old man, who made errors with calculations over 2 years, wore unmatched shoes and socks and tucked his jacket into his pants. Initially easygo-ing, he became stubborn and irritable. A religious awakening led him to spend hours in church; he argued with his wife and friends regarding his new religious ideas. His verbal output was fluent. A CT brain scan revealed mild atrophy, most marked in the right temporal lobe. A SPECT scan showed decreased cerebral blood flow in the anterior temporal regions, greater on the right. After he died, pathology was asymmetrically distributed, confirming the predominantly right temporal neuronal loss and gliosis.

In their summary of the clinical presentations, the authors commented that three out of five patients with the right-sided form of FTD had either increased religious ideas or heightened philosophical thinking.

A familial form of FTD was described by Lynch et al. (1994). Seven members of a family were identified with linkage to chromosome 17. The patients had person-ality and behavioral deterioration, in particular with signs of Parkinsonism, and initially release of abnormal behaviors, with disinhibition noted. Hyper-religiosity was noted in three. The behavior pattern was consistent with a frontal lobe dementia and, on pathological examination, all of the brains examined ($n = 6$) showed moderate to severe frontal and temporal lobe atrophy.

Saver and Rabin, in their discussion of the links between dementia and religious behaviors, make the point that in contrast to FTD, in Alzheimer's disease, along with the decline of many lifelong interests, religious concerns and practices also decline. They attempted to find a unified hypothesis of the neurology of religious experiences, largely relying on the epilepsy and dementia literature. They considered the primary substrate to be within the limbic system, and to be part of a distributed neural network marking events with either positive or negative valence. Acting as an intermediary between affects and cognitions, this cerebral system may mark experiences as either

depersonalized or derealized, as crucially important and self referent, as harmonious – indicative of a connection or unity between disparate elements, and as ecstatic – profoundly joyous.

This theory stands in contrast to, but alongside alternative hypotheses, which have emphasized the importance of the nondominant hemisphere in the modulation of holistic, nonverbal experiences, with the left hemisphere translating the experience of the nondominant hemisphere into an analytic and verbal version, which is inherently incomplete since the experience that is reported as in fact ineffable.

In the scheme of Savin and Raber, the perceptual and cognitive contents of the numinous experiences are seen as similar to those of ordinary mental experiences, but they are tagged by the limbic system as of profound importance united into a joyous unity. In their view, the descriptions of these experiences resemble those of ordinary experiences, but the distinctive feelings appended to them cannot be captured fully in words. They refer to "limbic markers," which can be named but cannot be communicated in their intensity, resulting in ineffability.

Disinhibition with social impropriety and offensive behavior are more a feature of the right temporal variant of FTD, and the authors concluded from their own observations that "the right hemisphere, at least the orbito-frontal and temporal aspects, may be necessary for mediation of socially appealing behaviour in humans." They go on to note, however, that since patients with epilepsy who have one or other of their temporal lobes removed at operation to help their epilepsy do not develop these severe behavioral syndromes, that the impaired behavior in their cases probably related to bilateral temporal lobe involvement.

Although their interpretation may be correct, that is about bilateral involvement, in keeping with the observations from the NHNN studies reported above, their observations about the effects of temporal lobectomy are not:

Patient TLE1: A female began to have seizures in her 20s. She had déjà vu auras, and then developed complex partial seizures and sometimes secondarily generalized attacks. Her EEG showed a right temporal lobe abnormality. She was mildly religious, but never went to church, and there was no psychiatric history before the operation, which was a right temporal lobectomy. The pathological study revealed amygdala and hippocampal sclerosis.

Immediately after her operation, the patient experienced the feeling of God's presence. She also had an auditory hallucination saying "its not time yet." She confessed that the experience was a revelation and that she believed that God was protecting her. She started attending church, often twice a week, and now goes to prayer meetings and meditation classes. She continues to feel the presence of God, and when strong, for example, in church, this is associated with feelings of euphoria and a racing pulse. She remains free of epileptic seizures.

Studies of Non-neurological Populations and Normals

These studies on patients with neurological disease are central to an unraveling of the neurobiology of religious experiences. Although the emerging behaviors can be

regarded as pathological, they have the same emotional tag and the same contents as those of the everyday experiences of millions of people who set about their religious observances in a socially appropriate fashion. It seems to be a fact that religious moments, inspirational and personally meaningful, are reported in a high percentage of normal people, perhaps up to 50 %. Most religious discourse relates to feelings about God or gods, and about beliefs in an afterlife or in reincarnation. In one survey in North America, 95 % of people confessed to a belief in God, and 71 % to a belief in the afterlife (Kroll and Sheehan 1989). However, in spite of the ubiquity of such sentiments in all social groups, there have been few studies that have addressed their potential neurological associations, or those of their behavioral counterparts, namely, the neural events and circuitry that may relate to behaviors of devotion.

Azari and her colleagues set out to test the hypothesis that religious experiences are not primarily emotional, that is related to a special feeling, the emotion in neurological terms being the feeling of the bodily responses as suggested by William James, but that they were cognitive-attributional (Aziri et al.). The areas of the brain activated by religious experiences induced in their volunteer subjects with religious recitation were cortical, especially the right dorsolateral prefrontal, the dorsomedial frontal, and the right medial parietal cortex (precuneus). No activation of the limbic structures was observed during the religious states, although in the happy state in the non-religious sample, the left amygdala was activated.

Although the states evoked in these volunteers are qualitatively different from those discussed in the neurological populations quoted above, the study is important in highlighting the cognitive components of religious states, and the role of the parietal cortex, especially of the nondominant hemisphere. The latter has direct connectivity with important structures in the limbic lobe, notably the parahippocampal gyrus and the cingulum, and as studies begin to unravel its function, the precuneus seems at least to be involved with autobiographical memory and states of self-awareness (Cavanna and Trimble 2006).

The work of Persinger and his collaborators, using volunteers and a special technique for stimulating the brain, covers two decades of scientific research into the links between the brain and religious experiences. He has also concerned himself with associations between symptoms and signs of temporal lobe disturbance and religiosity, and the role of the parietal cortex in such events (Persinger 2001). He notes that paranormal experiences are frequently associated with a sensed presence, involve the acquisition of information by a sense not regarded as normal, and contain distortions of time. Death or dissolution of the self are common themes, and embraced under the rubric "paranormal" are religious experiences, including sensing the presence of God, as well as other spiritual events such as haunts and alien abductions. Some of these experiences seem akin to those of the epilepsy patients discussed above. He refers to a continuum of "temporal lobe sensitivity" along which all human beings are distributed. Thus, using rating scales which measure temporal lobe experiences (temporal lobe sensitivity), his group reported that those individuals with the highest scores on

this scale also have more paranormal experiences, more frequent alpha rhythm over the temporal lobes on the EEG, and score higher on eccentric thinking and hypomania ratings using the Minnesota Multiphasic Personality Inventory, than those who score in the lower ranges for such temporal sensitivity. Using EEGs, he reported that temporal lobe transients (a wave of amplitude at least twice that of general activity – events he confusingly refers to as spikes) are significantly commoner in those with high temporal lobe sensitivity, and they correlated with religious belief or dogma clusters or "the feeling of presence" cluster on his measures.

Religious volunteers, as well as church attenders, score higher on temporal lobe symptomatology than the non-religious, and those reporting religious experiences were reported to be more likely to keep diaries, and enjoy poetry reading and writing (Persinger 1984a, b). All these data seem compatible with the epilepsy data reported above.

In another series of studies, Persinger stimulated the temporo-parietal regions of the brains of volunteers using very low frequency, weak magnetic fields, in the region of 1–5 μT in strength. This evoked in most normal people a feeling of a sensed presence, but the latter was more often elicited from those who had elevated scores on the measures of temporal lobe sensitivity. Interestingly, they were evoked most easily when applied for a period of 20 min after which "…a bilateral burst-firing pattern …was applied over the temporal poles for an additional 20 min" (Persinger 2001). While Persinger's results bear consideration, they also require much caution, not the least problem being their lack of replication by others. His equating of temporal lobe EEG changes in normal people with the underlying processes of epilepsy is highly questionable, as are some of his neurophysiological arguments. The neurophysiology is simply not that well worked out, and while the ability of microtesla stimulation of the brain to evoke paranormal experiences remains an intriguing observation, this does require independent verification.

However, his conclusion that religious experiences are evoked in normal brains by small electrical events in temporal lobe structures, that the parietal lobes are relevant to these experiences, and that bilaterality of stimulation may be a required feature of the response again is in keeping with the hypotheses which are emerging from the epilepsy studies.

Neurotheology

Neurotheology is a branch of neuroscience which has barely developed in recent years. However, there are now established University Departments that study the science of religion, and neuroscientists have taken up the challenge to investigate not only consciousness in its broadest context, but also emotional states, including those associated with religious feelings and practices. Neurotheology stands alongside other topics such as neuroaesthetics, as valid areas of neuroscience that can now be explored with, for example, brain stimulation and brain

imaging techniques. They represent neuroscience offshoots of what the biologist E O Wilson called sociobiology, namely, the systematic study of the biological basis of all forms of social behaviors, across the phylogeny, and including mankind. For him, religion was one such behavior which was universal, conferred genetic advantage, and, indeed helped drive evolutionary change (Wilson 1978).

This contribution reveals that there are, in reality, very few neurological disorders other than epilepsy that are associated with hyper-religiosity. The main one is FTD, and, taken with the epilepsy studies, the burden of the evidence favors bilateral limbic involvement but with an emphasis on changes in the right, nondominant hemisphere. The case cited above in this chapter also notes that removal of the right temporal lobe can release religious ideation and behaviors. This fact alone, plus the phenomenology of the described states, with the strong sense of presence of another in extracorporeal space, which is such a common denominator of patient experiences, implies involvement of other cerebral structures in such phenomena, and several authors have drawn attention to parietal lobe involvement. The volunteer studies strongly support this interpretation, especially the data from Azari and her colleagues and Persinger's group. The relationships of the parietal cortex to the body image, to the sense of self, and to states of self-consciousness require further studies, but they seem related to a circuitry which allows for the emergence of intense emotional religious experiences, bound with the activity of the limbic paleocortex. The dominance of sensory experiences, the predominance (but not exclusive) activation of the nondominant hemisphere, and the role of structures such as the precuneus, all help to explain the numinous, ineffable yet incorporated nature of religious epiphanies.

The forces of natural selection that shape our behaviors are discussed by nearly all contributors to this field. Mankind is, from an evolutionary point of view, young, but the human brain derives from eons of selective pressures. The genetic contributions to our behaviors, including religious behaviors, simply cannot be ignored. Indeed, there are several studies of the genetics of religiosity, the most convincing being that of identical twins reared apart, where the environmental influences on their behaviors will be different. Monozygotic twins, with ostensibly the same genetic apparatus, have been compared with dizygous twins, reared apart, on a variety of rating scales of religious beliefs and behaviors. It has been shown that monozygotic twins are more alike than dizygotic ones, and it has been calculated that some 50 % of the observed variance in religious experiences and beliefs is genetically rather than environmentally influenced (Waller et al. 1990). The molecular biologist Hamer (2004) has published his *book The God Gene*: *How Faith is Hardwired into Our Genes*, in which he claims not only that spirituality is adaptive from an evolutionary point of view, but that the gene(s) related to it are linked to those which regulate the neurotransmitters that control mood. There is no God-spot, as some lay commentators have naively tried to tie down some neuroscientists to claim, but *Homo sapiens* does bear the legacy of millions of years of mammalian neuronal tissue that has driven succeeding generations to thrive and

survive. The sapiens has arrived very late, but has arisen only upon development of the neurological substrates of our ancestors, revealing to us chthonic embedded behavioral patterns of the past. Religious experiences can only be interpreted in such a context, unless the discoveries of the last few hundred years of evolutionary biology and neuroscience are to be ignored.

Conclusions and Future Directions

Neuroscience, as a scientific discipline, must adopt a neutral stance to any investigations and subsequent results carried out in the area of brain and behavior, irrespective of the subject matter, and religious experiences and behaviors cannot be immune from such enquiries. At present, there are few hard and fast data to call upon in the area of neurotheology, and most of the writers recycle the same results from a small number of disparate investigators. In a recent review of the neuroscience of religious experience, McNamara (2009) sums up the anatomical findings as follows: "in summary, the circuit that mediates religiousness involves primarily limbic, temporal and frontal cortices on the right," (p. 129). Evidence is slowly accumulating, at least for some understanding of the cerebral associations of religious experiences. They are embodied, in the brain, and the key circuits involve the limbic circuitry and its associated pathways including the parietal lobes of, at least, the nondominant hemisphere. Future work will either support or refute these findings, but neurotheology, like its related discipline neuroaesthetics, will increase its field of influence over the next decade:

> Know then thyself, presume not God to scan,
> The proper study of mankind is man.[1]

Cross-References

- ▶ Cognition, Brain, and Religious Experience: A Critical Analysis
- ▶ Divine Understanding and the Divided Brain
- ▶ History of Neuroscience and Neuroethics: Introduction
- ▶ Human Brain Research and Ethics
- ▶ Neuroimaging Neuroethics: Introduction
- ▶ Neurotheology
- ▶ Toward a Neuroanthropology of Ethics: Introduction

[1]Alexander Pope. An Essay on Man, it continues:
He hangs between; in doubt to act or rest;
In doubt to dream himself a god, or beast;
In doubt his mind or body to prefer;
Born to die, and reas'ning but to err...
The pun on the word scan *is* intentional!

References

Azari, N. P., Nickel, J., Wunderlich, G., et al. (2001). Neural correlates of religious experience. *European Journal of Neuroscience, 13*, 1649–1652.

Bear, D. (1986). Behavioural changes in temporal lobe epilepsy: Conflict, confusion, challenge. In M. R. Trimble & T. Bolwig (Eds.), *Aspects of epilepsy and psychiatry* (pp. 19–29). Chichester: Wiley.

Cavanna, A., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129*(Pt 3), 564–583.

Csernansky, J., Leiderman, M. M., & Moses, J. A. (1990). Psychopathology and limbic epilepsy: Relationship to seizure variables and neuropsychological function. *Epilepsia, 31*, 275–280.

Dewhurst, K., & Beard, A. W. (1970). Sudden religious conversions in temporal lobe epilepsy. *British Journal of Psychiatry, 117*, 497–507.

Edwards-Lee, T., Miller, B., Benson, D. F., Cummings, J., & Russell, G. L. (1997). The temporal variant of frontotemporal dementia. *Brain, 120*, 1027–1040.

Foote-Smith, E., & Smith, T. J. (1996). Emanuel Swedenborg. *Epilepsia, 37*, 211–218.

Gastaut, H. (1956). La maladie de Vincent van Gogh. *Annales Médico Psychologiques, 114*, 196–238.

Hamer, D. (2004). *The God gene; How faith is hardwired into our genes*. New York: Doubleday.

Hood, R. W. (1975). The construction and preliminary validation of a measure of reported mystical experience. *Journal for the Scientific Study of Religion, 14*, 29–41.

Kannemoto, K. (2002). Post-ictal psychosis revisited. In M. R. Trimble & B. Schmitz (Eds.), *The neuropsychiatry of epilepsy* (pp. 117–134). Cambridge: Cambridge University Press.

Kroll, J., & Sheehan, W. (1989). Religious beliefs and practices amoung 52 psychiatric in-patients in Minnesota. *American Journal of Psychiatry, 146*, 67–72.

Lynch, T., Sano, K. S., Marder, K. L., Bell, L. L., et al. (1994). Clinical characteristics of a family with chromosome 17- link disinhibition-dementia Parkinsonism-amyotrophy complex. *Neurology, 1994*(44), 1878–1884.

Maudsley, H. (1869). Emanuel Swedenborg. *Journal of Mental Science, 15*, 169–198.

McNamara, P. (2009). *The neuroscience of religious experience*. Cambridge: Cambridge University Press.

Persinger, M. A. (1984a). Propensity to report paranormal experiences is correlated with temporal lobe signs. *Perceptual and Motor Skills, 59*, 583–586.

Persinger, M. A. (1984b). People who report religious experiences may also display enhanced temporal-lobe signs. *Perceptual and Motor Skills, 58*, 963–975.

Persinger, M. A. (2001). The neuropsychiatry of paranormal experiences. *The Journal of Neuropsychiatry and Clinical Neuroscience, 13*, 515–524.

Roberts, J. K. A., Robertson, M. M., & Trimble, M. R. (1982). The lateralizing significance of hypergraphia in temporal lobe epilepsy. *Journal of Neurology, Neurosurgery and Psychiatry, 45*, 131–138.

Saver, S. G., & Rabin, J. (1997). The neural substrate of religious experience. *The Journal of Neuropsychiatry and Clinical Neuroscience, 9*, 498–510.

Temkin, O. (1971). *The falling sickness* (2nd ed.). Baltimore: Johns Hopkins Press.

Trimble, M. R. (1991). *The psychoses of epilepsy*. New York: Raven.

Trimble, M. R. (2007a). *Biological psychiatry* (3rd ed.). Chichester: Wiley/Blackwell.

Trimble, M. R. (2007b). *The soul in the brain: The cerebral basis of language, art and belief*. Baltimore: Johns Hopkins Press.

Trimble, M. R., & Freeman, A. (2006). An investigation of religiosity and the Gastaut G Geschwind syndrome in patients with temporal lobe epilepsy. *Epilepsy & Behavior, 9*, 407–414.

Waller, N. G., Kojetin, A., Bouchard, T. J., Lykken, D. T., & Tellegen, A. (1990). Genetic and environmental influences on religious interests, attitudes and values. *Psychological Science, 1*, 138–142.

Waxmann, S., & Geschwind, N. (1974). Hypergraphia in temporal lobe epilepsy. *Neurology, 24*(7), 629–636.

Wilson, E. O. (1978). *On human nature*. Cambridge, MA: Harvard University Press.

Wuerfel, J., Krishnamoorthy, E. S., Brown, R. J., Lemieux, L., et al. (2004). Religiosity is associated with hippocampal but not amygdala volumes in patients with refractory epilepsy. *Journal of Neurology, Neurosurgery and Psychiatry, 75*, 640–642.

Cognition, Brain, and Religious Experience: A Critical Analysis

98

Aku Visala

Contents

Abstract

This chapter presents critical perspectives to neurotheology particularly in the light of cognitive science and philosophy. The main arguments revolve around the notion of "religious experience" and its usefulness in explaining religious phenomena. After providing a brief overview of the various forms of neurotheology and cognitive science of religion, it is argued that the study of religious experience might not be the best available starting point for the study of religion in general. This is mainly because the types of strong religious experiences that many neurotheologians study are unlikely to explain the prevalence of religious concepts and practices in general. The chapter also examines some more philosophical worries related to neurotheology. The main problem is the tendency to see the brain as the central explanatory factor of all our experiences and subsequent neglect of other levels of explanation (cognitive, everyday psychology). The final section highlights some of the positive

A. Visala
Department of Anthropology, University of Notre Dame, Notre Dame, IN, USA

Faculty of Theology, University of Helsinki, Helsinki, Finland
e-mail: aku.visala@helsinki.fi; avisala@nd.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_97,

contributions that neurotheology might make to the cognitive science of religion. These include focusing on the role of emotions in religious thinking and possibly challenging the deeply entrenched computational theory of mind in the cognitive science of religion.

Introduction

This chapter provides some critical perspectives to neurotheology particularly in the light of cognitive science and philosophy. The main issues revolve around the notion of "religious experience" and its usefulness in explaining religious phenomena. The chapter begins by looking very briefly at what kind of activities go under the headings of neurotheology and the cognitive science of religion. These preliminaries set the scene for the next part in which the different views of the centrality of religious experience in NT and cognitive science of religion are discussed. The third section addresses some general philosophical worries that have to do with NT. The final section highlights some of the positive contributions that NT might make to the cognitive science of religion.

Neurotheology and the Cognitive Science of Religion

The term "neurotheology" (NT) is somewhat vague, and it is unclear what kinds of activities it refers to. Normally, it has been used to refer to the neuroscientific studies of religious experiences, such the work of Andrew Newberg, Eugene D'Aquili, Michael Persinger, and several others, but there are at least two further kinds of research that could be considered as NT making three categories in total:

1. Neurotheology as the neuroscientific study of religious and spiritual experiences. Let us call this *neurotheology in the narrowest sense*.
2. Neurotheology as an activity in which neuroscientific results or data about the workings of the brain is religiously or theologically reflected and discussed. We can call this *neurotheology in a theological sense*.
3. Finally, neurotheology as a general program for scientific, religious, and theological study, as envisioned by Andrew Newberg (2010) and others. For them, neurotheology is a field of research combining science, religion, and theology in order to answer both scientific and religious questions. Newberg and some others (e.g., essays in Joseph 2003; Beauregard and O'Leary 2007) also draw religious conclusions from their research. Let us call this *neurotheology in the programmatic sense*.

NT understood in the narrow sense – as the neuroscientific study of religious and spiritual experiences – is strictly a scientific enterprise. The neuroscientist might be interested in studying, say, the experience of pain in connection with religious motivations or experiences, emotions linked to religious rituals or even altered states of consciousness apparent in some meditation practices just to mention a few topics. NT in a theological sense refers to theological or religious reflection of

neuroscientific results. Here the methods and assumptions are primarily theological, religious, or spiritual conceived in a broad sense rather than scientific (e.g., Watts 2002; Russell et al. 1999; Jeeves and Brown 2009). Oftentimes these two enterprises go together, especially when neuroscientists derive religious (or non-religious or anti-religious) conclusions from their results (e.g., Beauregard and O'Leary 2007).

The third referent for "neurotheology" is the proposed research program that attempts to include NT both in narrow and theological senses and expand their influence. According to Andrew Newberg, one of the architects of this program, "neurotheology refers to the field of study linking the neurosciences with religion and theology" (2010, p. 45). He goes on to emphasize that NT is not simply a scientific or religious enterprise but both, "Most importantly, neurotheology should be considered a two-way street with information flowing both from the neurosciences to the religious perspective as well as from a theological perspective to the neurosciences so that ultimately, both perspectives will potentially be augmented by the dialogue" (2010, p. 45). In its programmatic sense, NT constitutes an attempt to create a field of research in which scientists, theologians, and religious (and non-religious) people in general can reflect on the nature of religious phenomena and their relationship to our minds and brains. Newberg sees neuroscience in a very broad way as including cognitive neuroscience, psychiatry, psychology, and even social sciences such as sociology and anthropology.

Although NT in its programmatic sense seeks to encompass almost all scientific study of religion as well as its religious reflection, each activity can be distinguished from the others in terms of the goals of the people involved. The goal of the first type of activity is to produce scientific theories about the workings of the brain and how they contribute to (or correlate with) religious experiences. Bringing neuroscience to bear on religious phenomena might go further than just religious experience as, for example, in the work of Patrick McNamara (2009). Nevertheless, such pursuits are purely scientific activities and do not entail any explicit theological or (anti-)religious agenda (or at least it should not include such an agenda). In the second type of NT, the goal is the generation of some theological or more broadly speaking religious set of ideas that incorporates (or rejects) the scientific theories about religious experience and the brain. Here we have an activity that is best characterized as motivated and shaped by religious or philosophical rather than scientific interests.

Finally, there is NT in its programmatic sense that seeks to form a field that will supposedly have good scientific and religious effects. The ultimate goal of such a field would be a kind of Grand Unified Theory that combines both scientific theories and various religious and theological views about religious experiences and their relationship to our brains. Newberg calls this *megatheology* (e.g., Newberg 2010, pp. 64–65; D'Aquili and Newberg 1999, pp. 195–203). Such a megatheology would be a theology that potentially all religious people could accept and it would consist of neuroscientific explanations of mystical religious experiences and the more religious defense of their validity and usefulness.

Megatheology would be, if not orthodox in the traditional religious sense, at least more or less affirming the goods of spiritual or religious life. In what follows, when NT is referred to, the main meaning will be NT in the programmatic sense described here.

Compared to NT in its programmatic sense, *the cognitive science of religion* (henceforth, CSR) is a much more modest and narrowly defined enterprise. What the cognitive scientists of religion seek to do is to explain the cross-cultural recurrence of certain ideas and behaviors. Such behaviors and ideas include different kinds of rituals, beliefs about supernatural agents such as ghosts, spirits and gods, myths and divinely sanctioned morality. Instead of focusing on religious experience as a source of religion, the CSR looks at the cognitive underpinnings of religious thinking and behavior, namely, tendencies in human information processing that, given many kinds of cultural transmission events, create recurrent patterns. The central assumption here is that the human mind has some more or less natural information-processing tendencies and biases that are independent from specific cultural traditions. These tendencies and biases make some types of ideas easier to acquire, remember, and transmit than others. The prevalence of religious ideas suggests that they are particularly easy to adopt and transmit. In time, we would expect such ideas to become widely transmitted and shared, that is, to be incorporated into the prevailing culture. Several different lines of research provide evidence for this claim. Pascal Boyer and others (Boyer 1994, 2002) have argued that ideas that closely approximate our *intuitive assumptions* about the characteristics and causal properties of things in the world are easily acquired, remembered, and transmitted, compared to ideas that are *highly counterintuitive*, that is, ideas that require extensive learning and reflection. Boyer's claim is that religious ideas strike a balance between intuitive ideas and highly counterintuitive ideas and are, thus, attention-grabbing and memorable. Further, the greater the inferential potential of an idea, the more memorable it is. Inferential potential is understood here as an ability of an idea to generate inferences, predictions, and explanations pertinent to widely recurrent human concerns. Ideas about non-mundane agents (say, invisible spirits or ancestors) are both minimally counterintuitive (spirits are like humans but invisible) and have inferential potential (relevant for morality, misfortune, etc.).

Boyer's account has been supplemented by other theories that include a similar notion of cognitive naturalness. Scott Atran (2002) focuses more on the role of religious rituals as signals of commitment. Justin Barrett (2012) has argued that many religious ideas have a natural cognitive foundation that is manifested very early in human development. Developmental studies suggest, for instance, that children have a tendency to see many features of the natural world as designed or having a purpose and that this tendency is so strong it must be consciously overridden or else will persist into adulthood (Kelemen 2004). Such a tendency would make the notion of an intentional agent being behind the design of nature very natural. Further, some results suggest that certain divine attributes, such as "superknowledge" (having access to all relevant information about a particular situation) and immortality, are also natural for children, because children actually

understand all agents (not just gods) in these terms. Beings with minds are assumed to have complete knowledge, for example, until children learn otherwise. Finally, there are studies suggesting that belief in souls and afterlife are natural in the sense that they are supported by intuitions about minds, bodies, and death (Bering 2006; Bloom 2007).

The central argument of the CSR is, therefore, that religious ideas and behaviors are *natural* to beings with minds like ours. In order to avoid confusion, we need to explicate the technical notion of "naturalness" being employed here. As we have seen, "naturalness" refers to a certain kind of cognitive ease, but this comes in different varieties. Robert McCauley (2011) usefully distinguishes *maturational naturalness* from *practiced naturalness*. Practiced naturalness characterizes activities that require conscious training and cultural scaffolding. Riding a bike, driving a car, and writing are prime examples of practiced naturalness: When performed, they are automatic and easy, but they had to be learned through conscious effort and instruction. Maturationally natural activities, such as learning to speak one's native language and acquiring a set of basic moral feelings, are characterized by the same ease and automaticity but they arise without explicit instruction or learning. As a consequence, they seldom require special artifacts or teachers and emerge early in childhood. The main argument of CSR is at least some core cognitive processes associated with religious thought and behavior are maturationally natural. Compared with, say, science, chess, literacy, or bicycle riding, beliefs about gods and their moral nature, design in nature, and souls surviving death are largely natural in the maturational sense.

Religious Experience in Neurotheology and Cognitive Science

The category of "religious experience" has been notoriously difficult to define. In both philosophy and the study of religion, several different definitions have been offered, but no consensus has formed (Taves 2011). For the purposes of this chapter, it is enough to make a very rough distinction between religious experiences in a narrow sense and a broad sense. Religious experiences in the narrow sense refer here to experiences that are strongly non-mundane, that is, involve strong emotions and differ markedly from the experience of everyday life. Sudden visions, trance-like states, and other "altered states of consciousness" would qualify. William James' *Varieties of Religious Experience* (1985) is still a good source of vivid descriptions of experiences of this kind. Religious experiences more broadly construed can involve more mundane and long-term phenomena like the continuous feeling of God being present, and need not involve intense, punctual emotional arousal.

In terms of background, there is the debate between *perennialists* and *constructivists* about religious experiences (Taves 2011). For the perennialist, there is a shared, non-conceptual, or non-linguistic core to most religious experiences. Although some content and form of religious experiences depends on surrounding cultures and concepts, the core is independent from them. This core is usually understood in terms of pure experience or "pure consciousness." Against this, the

constructivist argues that there is no such thing as “pure experience” or consciousness, but instead all experience is mediated by concepts and language, which are in turn shaped by the surrounding culture. The debate is mainly about the correct approach to the study of religious experience: Should we explain what religious concepts look like with what religious experiences feel like or should we instead explain what religious experiences feel like in terms of religious concepts?

At first sight, the representative of the CSR would side with the constructivist and maintain that when we have explained the concepts and representations that are involved in religious experience, we have in fact explained the experience itself. Very roughly, the basic workings of the human mind explain why our concepts are what they are and they, in turn, shape our experiences. The representative of NT would be closer to the perennialist and maintain that there is an experiential core to religious experience that is not explainable in terms of cognition, information processing, and the cultural and linguistic context. In other words, there are some fundamental and cross-cultural truths about what, for example, unitary states of consciousness feel like and their explanation goes beyond explaining the concepts involved in describing the experience. For the neurotheologian, very roughly, the brain explains why our experiences have the characteristics that they do and the experiences, in turn, explain why we have the concepts that we do.

It seems fair to say that most of the work in NT has so far focused on the experiential and emotional aspects of religion. To be more specific, the main theories have to with “experiences of God” or some other type of “spiritual realities” and the practices that evoke these experiences (e.g., ritual, praying, meditation). For example, Michael Persinger (1987, 2003) has suggested that the “experience of God” is associated with deep microseizures in the temporal lobes. When Persinger stimulated magnetically certain parts of the temporal lobes, several subjects reported sensing “a presence of a sentient being.” The experiences were reported as making every aspect of the world feel profoundly meaningful, having a sense of unity, and being at peace with the world. Further, it might be the case that such “God experiences” can be triggered naturally by stress, anxiety, or by environmental conditioning linking the experiences to religious contexts. On the basis of the fact that when taken to the extreme, microseizures in the temporal lobes are connected with temporal lobe epilepsy (that manifests itself as psychic seizures without bodily convulsions), Persinger suggests that there is a connection between “mystical experiences” and temporal lobe epilepsy or other similar malfunctions. Ramachandran and others (Ramachandran and Blakeslee 1998) have also studied the connection between the temporal lobe and mystical experience by looking at persons with temporal lobe malfunctions or epilepsy. Their conclusion is that there is no clear evidence for religious stimuli being processed in the temporal lobe only. This suggests a more modest conclusion than that of Persinger’s. It is likely that there are some specific brain functions linked to religious experiences and that temporal lobe epilepsy sometimes involves intense “religious” experiences, but there is no clear connection between mystical experiences and the temporal lobe.

Another well-known NT theory is that of Eugene D'Aquili and Andrew Newberg (1999, 2002). Based on numerous brain-imaging studies of meditating or praying subjects, D'Aquili and Newberg claim that there is a hard-wired core experience they call *Absolute Unitary Being* (AUB). In the AUB state, "the subject loses awareness of discrete limited being and of the passage of time, and even experiences an obliteration of the self-other dichotomy" (1999, pp. 109–110). Unlike Persinger, D'Aquili and Newberg do not refer to any kind of brain pathologies here; for them, the AUB state is a product of a normally working brain in very specific circumstances. The neurobiological mechanisms undergirding these experiences are the sympathetic and parasympathetic nervous systems which rituals and prayers stimulate (for criticism, see, e.g., Atran 2003).

Now, the question is whether focusing on the neural correlates of religious experiences is fruitful for the study of religion. Newberg and other representatives of NT seem to think so. Contrary to this, cognitive scientists in general and the representatives of the CSR in particular usually think that focusing on religious experience (narrowly or broadly defined) is not very fruitful when explaining religion. There are several reasons for this.

First of all, most representatives of the CSR want to reject most traditional theories of religion in which religion is understood as having a specific essence or inner nature. Such theories of religion have often postulated a *sui generis* religious experience of, for example, the Holy (Rudolf Otto) or the Sacred (Mircea Eliade) as the core of religion. Since this experience and its object are *sui generis*, in a category of their own, the core of religion cannot be accessed outside the experience of the Holy or the Sacred. Such a view implies that there is a special "religious cognition" or some other psychological mechanism that is specific to religion. This is also what NT, at least partly, suggests: There are experiences that are unique to religion, and they have unique patterns of brain activation.

Against this view, the cognitive scientists maintain that there is neither a *sui generis* essence to religion or religious experience nor uniquely religious cognition (e.g., Boyer 1994). Instead, religion is considered to be a vague and diverse collection of different kinds of phenomena, such as rituals, stories, myths, moral feelings, and reasoning about supernatural agents and their powers. We need not assume, so the argument goes, that there is some unifying essence to all these phenomena. Instead, what we have is a collection of different kinds of thoughts and practices that utilize various non-religious cognitive mechanisms. Thus, we explain ritual in terms of our non-religious ways of representing actions or beliefs about souls and afterlife in terms of our non-religious intuitions about minds and bodies.

Another reason for the rejection of religious experience as an interesting topic of study is that the category of "experience" does not really function as an explanatory factor in cognitive science. This is because descriptions of experiences involve phenomenological notions and concepts, that is, concepts that describe what things "look like" to the person who is the subject of some particular experience. For the classical computationalist and functionalist paradigm of cognitive science, however, there are no such things as appearances or "seemings" in the workings of

the mind. The mind is (or can be treated as) a kind of computer transforming symbols. In other words, everything that the mind does can be explained in terms of computation of symbolic representations. Representations are, ultimately, coded in the brain in some form or another, and they involve nothing like "seemings" or "experiences." This view is often coupled with a view of *emotions* as highly specific programs that are designed by natural selection to produce certain kinds of behaviors (e.g., aversion, self-preservation). On this view, emotions are not defined or explained in terms of what they "feel" like, that is, what their phenomenal qualities are, but instead in terms of the informational input and resulting behavior. This approach is again typical of the computational/functionalist approach to the mind. The view amounts to a kind of "no experience" theory of the mind – would bracket out phenomenological notions, like "seemings" and feelings, as explanatory factors. Because of the prevalence of the computational/functionalist paradigm, in most (especially pre-2002) CSR work (e.g., Sperber 1996; Boyer 1994), "experience" of a religious subject is simply seen as the subject having certain kinds of beliefs. So having a religious experience is nothing more than having religious-type beliefs about the sources and causes of the events of one's surroundings.

More recently, some representatives of the CSR have tried to expand the theoretical resources of the CSR beyond this "no experience" view of early CSR and computational cognitive science. Ilkka Pyysiäinen has argued that religious experiences are more than simply beliefs: They involve emotions and evaluative judgments that guide actions (2001a, b). Pyysiäinen's theory of religious emotions draws directly from neuroscience and the work of Antonio Damasio (1996, 1999) and Joseph LeDoux (1996) in particular. Scott Atran (2002) has also argued that pure cognitive theories of religion are not enough and more attention should be directed to how religious beliefs and practices manage both positive and negative emotions in situations of "existential anxiety." Finally, the *Modes of Religion* theory of Harvey Whitehouse (2004) includes an explanation of the links between different kinds of rituals and their emotional intensity. Given these developments, it would be unfair to say that the CSR simply neglects the emotional or the volitional aspect of religion. Nevertheless, these theories are "cognitive" in the sense that they see religious experience in terms of information processing, not in terms of what these experiences feel like or what their neurobiological correlates are. Religious experiences are, very roughly, non-ordinary experiences interpreted in terms of religious concepts that elucidate strong emotions. This might be the place where NT has something to contribute to the CSR. I will return to this topic in the last section.

Finally, the third reason for rejecting religious experience as an explanation of religious phenomena has been the relative rarity of intense religious experiences like unitary or trance-like states. Again, nineteenth century psychologists of religion, such as William James, thought that religions stem from the experiences of sages and seers or "religious geniuses" like Jesus or Buddha. The great masses would then simply be impressed by the experiences of the geniuses and follow their insights. That James thought along these lines is unsurprising, since the romantic focus on the emotion and the heroic individual was the trend of the age.

The criticism later directed against such theories was that they are impotent in explaining the prevalence of religious behavior and the recurrent patterns in religious traditions. This is because religious experiences of the narrow kind exhibited by the "geniuses" are so rare.

Scott Atran (2003, p. 165) writes that "even if neurotheological speculations about the biological correlates of mystical experiences were true, there is no evidence that less extreme, more "routine," religious experiences have some characteristic brain-activation pattern." Atran then refers to a poll according to which approximately one third or one fourth of American and British subjects report having a religious experience, but only 2–3 % report having an intense "mystical" experience, such as feeling everything being one. So even if we are generous, the kinds of mystical experiences that representatives of NT such as Newberg and D'Aquili theorize about are quite rare among the religious. So to explain religion, it is not enough to explain the intense experiences of some "religious geniuses," because that would leave us without an explanation as to why religious people that have no such experiences (most religious believers) actually believe and understand what the geniuses are saying. Instead, what we need to explain is why people are ready to believe that non-mundane experiences have something to do with special realities, gods, spirits, and the like. In other words, why are people disposed to explain strange experiences in religious terms. Such an explanation, the argument concludes, needs something much more than just a neuroscientific explanation of narrowly defined religious experience. As Atran puts it, "mystical episodes may inspire new religions, but they do not make religion."

Atran summarizes the critique from cognitive science well, "despite intriguing findings concerning neurobiological correlates for certain types of intense religious experience, broader neurotheological interpretations of the findings are unwarranted. They involve speculation that not only strays way beyond the facts but crucially ignores or contradicts much recent work in cognitive and developmental psychology and cognitive anthropology" (2003, p. 165).

Additional Critical Perspectives

Representatives of NT often seem to think (or at least imply) that the constitution and workings of the brain are the only possible explanations of human behavior and thinking. It is rather typical of many representatives of NT to argue in the following way. Newberg and D'Aquili (1999, p. 45) write that "no matter what happens to us or what we do, there is a part of the brain that becomes activated. – Thus, the brain appears not only to react to everything that happens to us, but is eminently responsible for everything that we do or experience. In this way, studies of brain function help to show that . . . it is the brain by which all of our thoughts, feelings, and experiences are derived." The suggestion here seems to be that since all of our thoughts and behaviors involve brain events, they are best explained in terms of brain events. But this does not follow. Let us call this the "fallacy of micro-reduction."

The point can be made best by distinguishing singular causal explanations from *constitutional* explanations. We might wonder why a vase is broken on the floor in a pool of water. A singular causal explanation would give us the cause of this vase falling down from the table and breaking. One relevant explanation might be the strong wind that blew in from the open window. A constitutional explanation would answer a slightly different question, namely, why did the vase fell to the floor and broke instead of, say, flying up and staying in one piece. Now, the explanation would be given in terms of the physical structure of the vase and the floor and the general physical laws that govern their behavior. The point of this distinction is that the physical constitution of the thing being explained or the physical mechanism that produces the effect were explained is not necessarily the correct level of explanation. In other words, even if it the case that our thoughts are physical operations in our brains and that all our actions are, therefore, causally mediated by brain functions, it would not follow that all explanations of our thoughts and behaviors should be given in terms of what goes on at the micro-level of our brain (Craver 2007). Similarly, explanations in, say, biology and chemistry need not be given in terms of physics even though basic physical events and processes constituted all biological and chemical events and processes.

Atran (2002) coined the term "leapfrogging the mind" to pinpoint what he sees as the main problem of evolutionary explanations of human behavior. His argument is, very roughly, that when explaining what humans do, we have to explain how they believe and represent the world. Whatever biological or evolutionary causes human behavior might have, they have to be mediated through human cognitive systems. It is valid to look for an evolutionary explanation for some specific behavioral trait, but the explanation should not stop there. We also need an account of how evolutionary causes actually cause the behavior of humans in terms of the functioning of their cognitive system. So evolutionary explanations tend to skip an analytic level, that is, they leapfrog the mind and jump go straight from the evolutionary function of the behavior to contemporary human behavior without considering how human psychology works. The same kind of leapfrogging seems to be at work in NT in its programmatic form, in which the level that is skipped is the cognitive, the level of information processing. Behavior and thinking (the level of commonsense description) is explained by brain functioning (the level of neuroscience) without an account of the cognitive processes that mediate these levels.

The lesson here is that different levels of analysis need to be kept separate (at least at first). In our case, the relevant levels of analysis would be (1) the neurochemical and physiological mechanisms of the brain, (2) the information-processing mechanism of the mind, and (3) the everyday description of mental events, including descriptions of religious experiences and such. So even if there were some brain activation or a mechanism associated with everything we think and do, that does not mean that we could explain (or even should try to explain) all our behavior and thinking in terms of what happens in the brain.

There is a further point to be made about jumping or skipping levels. Cognitive theories of religious thinking suggest that religious thoughts are processed by cognitive mechanisms that also process non-religious thoughts. If this is true, when people

are reasoning about the intentions of God, gods, or spirits, they are using the same cognitive machinery that produces beliefs about the intentions ordinary, mundane human beings. Similarly, the same systems that we use to represent actions are used to make sense of religious ritual actions. This is simply a consequence of the more general point of the CSR maintaining that there is no "religious" cognition. If true, the prediction would be that on the level of brain function, there is nothing to distinguish ordinary religious thinking from ordinary non-religious thinking. On the level of neuroscience, therefore, there should be no difference between thinking about quantum mechanics, the nature of God as the First Cause, or the warrant for Darwinian notion of species. So no matter how hard we look at the ordinary religious brain that is not in a state of intense religious experience, we are unlikely to see anything that would distinguish it from a non-religious brain. Notice that what is being argued here is *not* that NT can explain nothing interesting or relevant about our experiences, behavior, or thought. Instead, the argument is simply that the neuroscientific level of analysis in which NT operates is not necessarily the best level of analysis in most cases of religious thinking and behavior.

Neuroscience in general and NT studies in particular often suffer from a problem that somewhat resembles the issue of micro-reduction and level skipping. At least on the surface level, neurotheological texts are rife with talk in which all sorts of activities are attributed to the brain. Brains, we are told, think, perceive, and have trance-like states and absolute unitary experiences. Not only does the brain "experience" the world, but different parts of the brain might "have different experiences" or "perceive the world in different ways." This may all be just a sloppy use of language, but the prevalence of the talk is indicative of a view according to which the human brain is the agent of human action. This, of course, would fit in very well with the micro-reduction that was just mentioned. If the brain is the cause of human actions and thoughts, it is the brain that is the agent of human behavior, not the person herself. The "person" or the "self" is only a construct of the brain, not the cause or the subject of actions and thoughts.

As with micro-reduction, there is a problem here. In their book *The Philosophical Foundations of Neuroscience* (2003), Maxwell Bennett and Peter Hacker argue that it is a mistake to attribute the properties of the whole to the parts of the whole. They call this the *mereological fallacy*. Let us consider the act of writing an essay on a laptop. It is not the computer motherboard or the processor alone that does the word processing, but the whole computer with all its functioning parts. The computer would be unable to perform its task if the motherboard or the processor was missing, but this does not mean that they are the subjects of word processing. So it is a mistake to attribute the action of word processing to the processor or to the motherboard. Furthermore, it is not the hands of the writer that write the essay, but the person. The hands only press the keyboard. Again, the person would be unable to write the essay without his hands, but this does not mean that we should attribute "writing" to hands themselves. So the proper subject for the act of writing is not the hand of the person, it is the person that is writing.

Now, Bennett and Hacker argue that, by the same token, it is not the brain that is the proper subject of "perceiving," "experiencing," or "thinking," it is the person.

Despite the fact that there might not be a unified "self" or a "person" in some deep sense (Damasio 2012), we can attribute complex actions and thoughts only to beings that have self-regulated movement and are a part of some kind of linguistic community. In other words, it makes no sense to say that a being is thinking about its mother, if the being in question is incapable of intellectual life and expressing that life in some way or another (bodily action, language and so on). The point that Bennett and Hacker make here is that our everyday mental language does not really apply to things like brains, it only applies to beings with acting bodies and the capacity of complex linguistic expression. Thus, instead of simply attributing our actions to the brain, we need an account of how the brain is related to the way in which *we* "perceive," "experience," and "think."

Possible Contributions from Neurotheology

This section highlights several ways in which NT might contribute to the CSR. The main theme is the aforementioned tendency of cognitive theories of religion to focus on beliefs and concepts. In its modest forms, NT could provide evidence for and remind the cognitive scientist that there are many conscious, volitional, emotional, and experiential aspects in religious traditions.

The functionalist/computational paradigm is rather deeply entrenched in both the CSR (Boyer 1994, 2002) and in evolutionary psychology (Pinker 1997). In addition to understanding metal operations in terms of computation, the paradigm entails that the actual physical basis of thought is not really important. Since mental states are defined in terms of functional roles, the physical basis is "multiply realizable." That is to say, the same mental operation can be performed by different physical stuff: computers, human minds, or even large enough collections and organizations of cans of beer. The type of matter (nerve cells, silicon, etc.) does not make a difference with respect to thinking, only the way in which the matter is organized to perform is significant. Because of "multiple realization," the levels of information processing and its physical workings are independent from each other. It follows that the advocate of the functionalist/computational paradigm is unlikely to be interested in the way in which the brain actually realizes the different information-processing functions. As we have seen, this approach also directs the focus to belief and representation and away from other mental phenomena, like emotion, consciousness, volition, and imagination that are more difficult to conceptualize in computational terms. Although CSR admits that our cognitive processes are mostly "hot" instead of "cold," that is, emotional and volitional cognition often overrides abstract or more conceptual cognition, there is a tendency to focus on religious thinking in abstract, conceptual terms only.

By contrast, several neuroscientific studies point to many different kinds of emotional, physical, and volitional effects of religious and spiritual practices and experiences (e.g., Newberg and Waldman 2009). These results suggest that "conceptual" cognition is not separate (or wholly separate) from the emotional aspects of cognition. Indeed, some neuroscientists have argued that rational thinking is

impossible without emotional input (Damasio 1996). NT studies can function as reminders that there might be interesting religious phenomena that are difficult to study without taking into account the broader affective as well as cognitive experiences of religious subjects. This broader understanding might be particularly important in studying intense religious experiences.

NT could also contribute by producing evidence for or against certain assumptions that cognitive and biological theories make about how the human mind works. Cognitive scientists often postulate various information-processing modules and systems, such as the hypersensitive agency detection device (HADD), different mechanisms of social cognition, or the Theory of Mind mechanism. The assumption is that neural correlates for such mechanisms could eventually be found or at least their workings explicated in neuroscientific terms. Even if a cognitive scientist were strongly committed to the functionalist/computational paradigm, there would still be some kind of neuroscientific story to be told about the realization of specific information-processing modules. Conversely, let us say that no neuroscientifically valid story of the workings of some cognitive mechanism can be found. Even for the most hardened computationalist, this outcome would constitute at least *prima facie* evidence against the existence of such a mechanism as an independent entity.

Conclusion and Future Directions

In conclusion, let me highlight one critical issue that I have not explicitly discussed above. Behind neuroscience and cognitive science, the issue of consciousness and the nature of the "self" loom large (see, e.g., the brief analysis in Taves 2011). What is the relationship of cognitive and neuroscientific explanations to our conscious reasoning and to what extent these approaches account for the phenomenal qualities of our conscious experiences? Both NT and CSR (wisely) do not take strong views on this question. However, it might be that the issue of consciousness is not so easily bracketed. On the functionalist/computational theory of mind, phenomenal conscious states are either reducible to certain kinds of representations or they are epiphenomenal, that is, side effects of something else and have no effects by themselves. Further, many neuroscientists seem to think that since we cannot find consciousness or selves in the brain, there are no such things and everything humans do and think is explained in terms of the physical operations of the brain. Now, the computational theory and the cruder forms of neuroscientific reductionism are not without their critics. The computational view has been subjected to sustained criticism from both neuroscience and philosophy (e.g., Searle 1989; Fodor 2000). Similarly, the excessive use of neuroscience has been strongly criticized by, e.g., Raymond Tallis and others (Tallis 2011). The main critical point, it seems, is that both neuroscience and computational cognitive science are unable to make sense or take into account the phenomenal aspects of consciousness and the "aboutness," intentionality, of mental states and are, thus, ill equipped to deal with explicitly conscious aspects of human behavior and thought.

To fend off possible misunderstandings, let me also emphasize what I did *not* argue in this chapter. The argument was *not* that all forms of neurotheology should be rejected. The aforementioned arguments against approaches that focus on religious experiences need not entail a total rejection of NT. What they do entail (if correct) is that a neuroscientific study of religious experiences might not be the best starting point of a general theory of religious thought and behavior. Furthermore, the aforementioned arguments apply to some forms of NT only to some extent or none at all. NT in its programmatic form is clearly the main target of these criticisms. If one is a representative of NT in its narrow sense, the criticisms might not apply at all. If one is sufficiently modest and interested in studying the neural correlates of, say, experiences associated with trance-like states, one need not make sweeping claims about there being an experiential essence to such experiences cross-culturally or that one could develop a theory of religion on the basis of such a research. In addition, NT in its narrow sense need not entail that the experiences that neuroscientists study are not conceptually structured or conditioned. Again, the argument is not that NT is useless, but that it needs to be more modest in its claims.

As for the future of the scientific study of religion, it seems to me that we need neuroscience and cognitive science as well as explicitly evolutionary approaches. Religion is a multi-faceted phenomenon that has its roots in the way we human beings are and how we have evolved. This includes our brains as well as our minds. There is a desperate search for both conceptual tools and empirical data that would help us evaluate current theories and relate them to one another. Here the future scientific and philosophical developments around two major interfaces, the brain/cognition (neuroscience/psychology) and cognition/conscious thinking, will most likely provide useful insights into religion as well.

Cross-References

▶ Divine Understanding and the Divided Brain
▶ Explanation and Levels in Cognitive Neuroscience
▶ Model-Based Religious Reasoning: Mapping the Unseen to the Seen
▶ Neurotheological Eudaimonia
▶ Neurotheology
▶ Realization, Reduction, and Emergence: How Things Like Minds Relate to Things Like Brains

References

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion*. New York: Oxford University Press.

Atran, S. (2003). The neuropsychology of religion. In R. Joseph (Ed.), *NeuroTheology: Brain, science, spirituality, religious experience* (pp. 147–166). San Jose: University Press.

Barrett, J. (2012). *Born believers: The science of Children's religious belief*. New York: Free Press.
Beauregard, M., & O'Leary, D. (2007). *The spiritual brain: A neuroscientist's case for the existence of the soul*. New York: HarperCollins.
Bennett, M., & Hacker, P. (2003). *The philosophical foundations of neuroscience*. Oxford: Blackwell.
Bering, J. (2006). The folk psychology of souls. *Behavioral and Brain Sciences, 29*, 453–462.
Bloom, P. (2007). Religion is natural. *Developmental Science, 10*, 147–151.
Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion*. Berkeley: University of California Press.
Boyer, P. (2002). *Religion explained: The human instinct that fashion gods, spirits and ancestors*. London: Vintage.
Craver, C. (2007). *Explaining the brain: The mosaic unity of neuroscience*. Oxford: Oxford University Press.
D'Aquili, E., & Newberg, A. (1999). *The mystical mind: Probing the biology of religious experience*. Minneapolis: Fortress Press.
Damasio, A. (1996). *Descartes' error: Emotion, reason and the human brain*. London: Papermac.
Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.
Damasio, A. (2012). *The self comes to mind: Constructing the conscious brain*. London: Vintage.
Fodor, J. (2000). *The mind does not work that way: The scope and limits of computational psychology*. Cambridge: The MIT Press.
James, W. (1985). *Varieties of religious experience*. Harvard: Harvard University Press.
Jeeves, M., & Brown, W. (2009). *Neuroscience, psychology and religion: Illusion, delusions and realities about human nature*. West Conshohocken: Templeton Foundation Press.
Joseph, R. (Ed.). (2003). *NeuroTheology: Brain, science, spirituality, religious experience*. San Jose: University Press.
Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and reason in nature. *Psychological Science, 15*, 295–301.
LeDoux, J. (1996). *The emotional brain: The mystical underpinnings of emotional life*. New York: Simon & Schuster.
McCauley, R. (2011). *Why religion is natural and science is not*. New York: Oxford University Press.
McNamara, P. (2009). *The neuroscience of religious experience*. Cambridge: Cambridge University Press.
Newberg, A. (2010). *Principles of neurotheology*. Farnham: Ashgate.
Newberg, A., & D'Aquili, E. (1999). *The mystical mind: Probing the biology of religious experience*. Minneapolis: Fortress Press.
Newberg, A., & D'Aquili, E. (2002). *Why god won't go away: Brain science and the biology of belief*. New York: Ballantine Books.
Newberg, A., & Waldman, R. (2009). *Hod God changes your brain: Breakthough findings from a leading neuroscientist*. New York: Ballantine Books.
Persinger, M. (1987). *Neuropsychological bases of god beliefs*. New York: Praeger.
Persinger, M. (2003). The temporal lobe: The biological basis of God-experience. In R. Joseph (Ed.), *NeuroTheology: Brain, science, spirituality, religious experience* (pp. 279–292). San Jose: University Press.
Pinker, S. (1997). *How the mind works*. New York: Norton.
Pyysiäinen, I. (2001a). *How religion works: Towards a new cognitive science of religion*. Leiden: Brill.
Pyysiäinen, I. (2001b). Cognition, emotion and religious experience. In J. Andresen (Ed.), *Religion in mind: Cognitive perspectives on religious belief, ritual and experience* (pp. 70–93). Cambridge: Cambridge University Press.
Ramachandran, V. S., & Blakeslee, S. (1998). *Phantoms in the brain: Human nature and the architecture of the mind*. London: Fourth Estate.

Russell, R., et al. (1999). *Neuroscience and the person: Scientific perspectives on divine action*. Vatican City: Vatical Observatory.
Searle, J. (1989). *Minds, brains, and science*. Harmondsworth: Penguin.
Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Oxford: Blackwell.
Tallis, R. (2011). *Aping mankind: Neuromania, darwinitis and the misrepresentation of humanity*. Durham: Acumen.
Taves, A. (2011). *Religious experience reconsidered: A building block approach to the study of religion and other special things*. Princeton: Princeton University Press.
Watts, F. (2002). *Theology and psychology*. Aldershot: Ashgate.
Whitehouse, H. (2004). *Modes of religiosity: A cognitive theory of religious transmission*. Walnut Creek: Altamira Press.

Model-Based Religious Reasoning: Mapping the Unseen to the Seen

99

Adam Green

Contents

Abstract

Contrary to what might be commonly believed, there is a family resemblance between model-based reasoning in the sciences and popular religious reasoning. Model-based reasoning factors into popular religious reasoning in large part due to the role that model-based reasoning plays in social cognition generally, and the role that social cognition in turn plays in thinking about extra-natural agents. Specifically, popular religious reasoning can be considered as an instance of the human tendency to isolate the structural features of peculiar events and to match these features to some overall view of the world. The procedure is not unlike the use of a particular map while hiking, with certain features that stand out in the terrain matched to symbols on the map. Just as one can interpret one's environment in terms of a preexisting map (a "top-down" approach), however, it is also possible to abstract the structural features of one's environment to create a new map (a "bottom-up" approach). This chapter suggests that one can think of natural theology as a form of bottom-up model-based reasoning, one that mirrors the top-down model-based reasoning common in popular religious reasoning.

A. Green
Department of Philosophy, Azusa Pacific University, Azusa, CA, USA
e-mail: Greenab3@gmail.com

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_98,

Introduction

One does not have to be an atheist for much popular religious reasoning to seem peculiar or even intrinsically irrational. From a perspective formed by a scientific outlook, popular religious explanations, such as that cancer is caused by witchcraft, that socioeconomic privileges reflect karma, or that a tsunami is an act of divine judgment, seem well beyond the warrant of the available evidence and harmful to true understanding. Moreover, popular religious explanations for cases of fortune and misfortune, origin and purpose, can seem far removed from the refined use of reason one finds in natural theology as well as the rigors of scientific explanation. Contrary to initial appearances, however, this chapter aims to show that the pattern of thinking involved in popular religious reasoning bears at least a family resemblance to model-based reasoning in the sciences as well as to natural theology.

Model-Based Reasoning

To explicate model-based reasoning in general, this section follows an approach in the spirit of Ronald Giere's model-based philosophy of science, especially as developed in Giere (2006). Rather than Giere's claim that maps and models are merely analogous, however, and that maps are physical objects whereas models are abstract entities, this section proposes that maps can be instantiated inside or outside the mind, and are one species of a more general class that one can usefully identify as that of "models."

Consider the example of using a map while on a hike. What exactly makes the piece of paper in one's hand successful as a "map," able to help one to navigate reliably and safely? One possible answer would be that the piece of paper is a map, because it represents the terrain. Unlike the way a king's seal represents the approval of a king, however, the map represents the terrain due to the fact that it resembles the terrain. Nevertheless, resemblance alone does not suffice to account fully for what makes something a map, since all pieces of paper resemble the terrain in at least some sense, such as both being physical entities. Moreover, there will be many things the map resembles more than the terrain, such as other copies of the same map.

To a first approximation, what makes the piece of paper a map is that it resembles the terrain due to the manipulation of the spatial relationships of items on the paper so as to track the spatial relationships of certain features of the terrain. One can have trivial and even accidental instances of such a resemblance. For instance, one's living room furniture can happen to be arranged in a way that corresponds to the layout of Trafalgar Square. In the case of maps, however, the resemblance between the map and what it is a map of is *intentionally* proportional, with the selected features and their proportionality being tied to a set of intended uses to which the map can be put. The degree of resemblance does not have to be too high for a piece of paper with symbols on it to function as a map, and indeed too much detail can be unhelpful. Rather, the resemblance has to be sufficient for the map to be used for its intended purposes, such as navigation on a hike.

In order to make use of a map to reason about one's future courses of action, however, one also needs to know how to match what one can observe in one's surroundings to part of the map, thereby unlocking further information from the map about various other matters in one's surroundings that one cannot observe directly. For example, a conventional symbol on a map may assist in discovering whether a building in the distance is a ranger station or a private residence. The map may also provide insight by placing what is observable into a larger context that cannot be observed directly, such as when one wants to discover exactly how much farther one has to hike before making it back to civilization. Moreover, it should be noted that, beside particular facts about local geography, consulting the map may also produce a Gestalt effect, allowing one to organize mentally what is observed on the terrain in the light of the structure of the map.

The features of maps and map-based reasoning identified here apply much more broadly. Translating the relational structure of something into the relational structure of a different thing need not take the form of translating spatial relationships in the target into the spatial relationships of a representational proxy. A bar graph depicting the proportion of redheads in the population does not resemble the way redheads are spatially located in the world vis-à-vis others, not least because there is no box-like pile of redheads anywhere in the world. Nevertheless, the spatial structure of the representation, rather than convention or causal history alone, communicates the intended information by resembling, albeit by means of an abstraction, some selected features of the world. The broader category that includes maps, without being limited to them, may be called "models." Models in general take relational structures in one domain and use them as a way to gain cognitive understanding of equivalent relational structures in another domain.

Such a definition is, of course, extremely broad and mention should be made in passing of the many important differences between various kinds of model-based reasoning. First, for example, it may be the case that reasoning with models about complex relationships in a visual modality is slower than using more abstractly represented models, perhaps due to cognitive noise created by extraneous visual detail (cf. Knauff et al. 2003; Knauff and May 2006). Second, using an external representation of a model may change the nature of the internal operations employed by model-based reasoning. What may be the same model may engage reasoning processes that are quite different depending on what cognitive tools are available. See, for instance, the classic case study of reasoning in navigation in Hutchins (1995). Nevertheless, such differences do not exclude an underlying family resemblance between the various kinds of models and model-based reasoning that we use (cf. the general theory presented in Goodwin and Johnson-Laird 2005 which is supposed to apply cross-modally).

Within this broad definition, it is quickly evident that models play an important role in scientific reasoning and explanation (cf. Frigg and Hartmann 2012; Giere 1988, 2006; Thagard 2012). Consider the following generic example. Suppose that data points are plotted on a two-dimensional coordinate system. The x and y axes of the coordinate system represent, in graphical form, two experimentally observable and measurable characteristics of the target(s) of one's research, characteristics that

may not themselves be spatial, such as a heat capacity and a temperature. One may think of the plot of points as an instance of a model such that the relation of the points plotted in the coordinate system stands for a relation of another, non-spatial type in the target domain. Typically, outliers in the collected data are then discounted and a curve is fitted through the critical concentrations of data on the coordinate system. The result is a model of the original data that simplifies what is taken to be relevant so as to make it more usable and, in particular, more usable for theorizing. Although there are deep philosophical questions about the nature of this process, such work is done in the hopeful expectation that the idealized model of the data approximates some underlying causal structure of the world.

Modeling, however, can also be used in science in a slightly different way, namely, in comparing experimental data with some theory that is already known, either for the purpose of interpreting the data or testing the theory. In fact, there is a more restrictive use of the term “model,” tied to formal methods in logic and computation, that would seem to apply most naturally to models of this sort (cf. Hodges 1997; Thagard 2012). In contrast to the previous example, one already has a theory about some deep underlying causal structure of the world prior to conducting the experiment and one wants to relate the data obtained to theoretical entities that are not themselves directly observable. What is needed is the ability to map the data about observable properties, plotted in a curve on a two-dimensional coordinate system, onto relationships between purported theoretical entities. To achieve this goal, one will often need to develop a representation of the theory’s structural properties or translation schemes for those properties that are amenable to comparing the structure of one’s theory and the structure of one’s data.

The latter kind of modeling is important for a number of reasons. First, even when theories take the form of universal rules, they may have adjustable parameters or various kinds of idealizations that require the mediation of a model that maps them onto specific contexts so that their predictions can be compared to experimental data. Theories of physics, for instance, are often described in terms of idealizations, such as frictionless planes or closed two-body systems, that invariably have to be related to the conditions of an actual experiment, normally by means of some kind of model, in order to assess the degree of resemblance of the theory to relevant features of the world. An example is the way that particle physicists have to model their own experimental apparatus in great detail in order to relate the effects of proposed universal laws to the details of an actual experiment, with varying kinds of signals, “dead space” (unresponsive parts of the detector), and so on. Second, the theory can be richer in information than one’s available data, so that a partial, simplified representation of the relevant theoretical domain is needed in order to make a meaningful or practical comparison, as in the case of the hiker who needs a map to give directions for walking and not, for example, details of the underlying geology of the landscape. Third, theories themselves may often be in the form of models that capture some but not all of the characteristics of an actual physical system, such as a billiard ball model of a gas or a computer model of the brain.

Finally, it should be noted that scientific modeling can also be effective in theoretical and practical terms even when the underlying physical causes remain opaque. Consider, for example, the transition in the nineteenth century from the miasma theory to the waterborne theory of cholera (cf. Johnson 2006). In the 1850s, no one could explain the agency that produced cholera, why the disease manifested itself with the symptoms that it did, or predict where it would occur. Nonetheless, it was possible to compare one theory, that cholera was caused in some unknown way by exposure to the atmosphere in certain poor parts of a city, with the alternative championed by Dr. James Snow, that the disease was somehow caused by tainted water. What Snow did was to construct a map of the deaths due to cholera in relation to particular sources of water. The concentrations of the disease around particular water supplies were sufficient to allow Snow to demonstrate the efficacy of his theory over its competitor, even though both theories lacked any account of the underlying physical cause of the correlations between the water and the disease, namely, the spreading of the bacterium *Vibrio cholerae*.

Models, then, clearly enter into scientific reasoning in a wide variety of ways: as idealizations of experimental data that may also indicate underlying causal structures of the world; as means to compare proposed theoretical entities with experimental data; and as ways of uncovering correlations between empirical phenomena even when the underlying causes remain opaque. In all these and other ways in which models and model-based reasoning enter scientific practice, they enable new insights by mapping structural features in one domain onto another domain, even in the absence of a wholly satisfactory account of the underlying causes of the observed phenomena.

Social Reasoning and Popular Religion

Besides specialized applications in the sciences, model-based reasoning is arguably ubiquitous in everyday life and especially in the social domain. Overt references to models and model-based reasoning in the mind-reading discussion are relatively rare at present, exceptions being Peter Godfrey-Smith and Heidi Maibom in their accounts of social cognition (Godfrey-Smith 2005; Maibom 2009). Nevertheless, as will be shown, describing various kinds of social cognition in terms of model-based reasoning is quite intuitive. Moreover, much research in the cognitive science of religion has converged on the thesis that a great deal of what goes on in religious thinking is subserved by everyday social cognition redeployed to extra-natural agents (Cf. Atran 2002; Barrett 2004; Bering 2011; Boyer 2001; Guthrie 1993; McCauley and Lawson 2002). Hence if model-based reasoning is important for social cognition, it is also important for neurotheology. More specifically, model-based reasoning is plausibly central to the crucial transition between the “mere” detection of some purported agency, whether human or divine, and an understanding of the inferred agent’s perspective and possible future behavior.

How then does model-based reasoning enable this transition to more complex kinds of theological reasoning than agency detection? As a more immediately familiar example of how such model-based reasoning works in everyday life, suppose that a child called Suzie steals the favorite action figure of her sibling, Tommy. The next morning Suzie does not find the stolen action figure where she had hidden it, but instead sees her favorite comic book ripped neatly in two. Suzie quickly comes to the conclusion that the ripping of the comic book was an intentional, retaliatory action on the part of Tommy, even though she did not see Tommy rip her book and was not told by anyone that he did so.

How does Suzie come to this conclusion? It would be psychologically unrealistic to attribute to Suzie the possession of a set of universal rules of human psychology that are so detailed and encompassing as to contain information about comic books and their demise. A much more plausible gloss is as follows. Suzie has certain expectations about how the world works. Against the backdrop of these expectations, the fate of the comic book stands out as a singular and unnatural state of affairs best understood in terms of intentional action rather than the standard, non-agential way in which events usually transpire. Having represented the ripping of the comic book as intentional, Suzie matches this action against her relationships with possible perpetrators. The action does not fit her normal expectations for how Tommy behaves, but it is easier for her to attribute the ripping of the comic book to an intentional action of Tommy than for her to see it as an accident or the action of someone else. Moreover, given the special circumstance, namely, that she has just stolen Tommy's favorite action figure, and that the action figure is now missing from her possession, a possible indication that Tommy has discovered the theft and recovered his property, she concludes that he has a possible motive for retaliation. She therefore concludes that Tommy ripped her comic book.

Suzie can be usefully thought of as engaging in model-based reasoning at each of the levels covered in the previous section. She identifies an anomaly in her environment, perhaps against the backdrop of a general model of the world and how it should unfold over time when not interfered with by agents. She isolates the relevant features of the anomaly and uses those features to fit the ripping of the comic book into the "effect" slot of her intentional action schema. She then looks to fill out that schema either by consulting a standing model of the psychologies of her family members or by enacting a new model of them through imagining herself in their respective positions. She consults the recent history of her relationships, which allows her to bridge the gap between her model of what has happened to her comic book and her general psychological model of her brother. This allows her to form a new psychological model updated with the special circumstances surrounding the stolen figurine that fits nicely with the tearing of the comic book and gives Suzie a useful explanation of what has happened.

No doubt there are differences between what Suzie does and what scientists do when they engage in model-based reasoning. For one thing, Suzie is probably more likely to apply her models as heuristics instead of through explicitly considered and endorsed discursive reasoning. Suzie recognizes a fit between what are, in fact, models, but she is not likely to be fully aware at a meta-level of what she is doing

when she reasons this way. The claim is not, of course, that heuristics are not also used in science. The work of Herbert Simon and Gerd Gigerenzer, for example, gives one reason to suspect that heuristics will generally be involved in mature, specialized cognition rather than being a mere "vice" of untrained and unreflective reasoning (cf. Simon 1977; Gigerenzer 2008). Rather, Suzie's social cognition is more likely to be what Bob McCauley calls "maturationally natural," like learning to walk or speak, whereas scientific model-based reasoning requires "practiced naturalness," like learning to ride a bike or play the piano (McCauley 2011).

Despite such differences, however, a family resemblance remains between scientific model-based reasoning and what Suzie does. Both processes seek to identify relational structures and map them against other relational structures in an effort to explain some aspect of the world. Moreover, these processes rely on something that is apparently like modeling both at the level of conceptualizing the *explanandum* and at the level of bridging the gap between the *explanandum* and candidate *explanantia*. Furthermore, in both cases, the key to successful model-based reasoning is to develop a representation of a target phenomenon that helps connect that target to a more general picture of the world.

The pattern of steps that takes place in social scenarios such as these also seems similar to much popular religious reasoning. Whatever a religious person thinks of religious experience, she should be amenable to the idea that religious explanations that are not rooted in qualitatively unique experiences are processed through a redeployment of whatever mental machinery is normally involved in social explanations. Nevertheless, setting aside cases that are experienced by the subject in some way that is especially conducive to the attribution of extra-natural agency, for what reason would such agency normally be invoked as an explanation?

One possibility is that the reasoning is similar to how Suzie comes to the conclusion that Tommy has ripped up her comic book. Some event strikes a person as peculiar, a deviation from the normal course of things that suggests that it is the effect of some intentional action: for example, the event might promote or contravene one's own interests more than one would expect to be true of a genuinely random event; or perhaps the peculiar features in question resemble the intentional design of human artifacts. If the cast of possible natural agents does not account for the putative intentional action, then one is faced with a disjunction. Either, contrary to appearances, the event is a random fluctuation in the normal course of things, or an extra-natural agent is involved. If one's metaphysics includes extra-natural agents who can intervene in the world, and their motives and interests can be matched to the event, explaining the event in terms of such agency is as understandable as Suzie inferring an intentional explanation for what happened to her comic book rather than her supposing that the book spontaneously divided into two apparently ripped pieces.

The foregoing account should not be taken to imply that the peculiar events that elicit this kind of religious reasoning must be taken to be rare, or that the reasoning involved is always as uncontroversial as it is in the case of Suzie's comic book. Depending on one's cultural context, temperament, and perhaps character, one might ascribe events to extra-natural agency quite promiscuously.

Rather, the key idea is that the events in question are taken to stand out as having a structure that deviates from the normal way of things in the absence of intentional action. Consider, for example, the following anecdote recorded by the anthropologist E. E. Evans-Pritchard, concerning a boy of the Zande people, who see witchcraft as permeating life.

> A boy knocked his foot against a small stump of wood in the centre of a bush path, a frequent happening in Africa, and suffered pain and inconvenience in consequence. Owing to its position on his toe it was impossible to keep the cut free from dirt and it began to fester. He declared that witchcraft had made him knock his foot against the stump. . .I told the boy that he had knocked his foot against the stump of wood because he had been careless, and that witchcraft had not placed it in the path, for it had grown there naturally. He agreed that witchcraft had nothing to do with the stump of wood being in his path but added that he had kept his eyes open for stumps, as indeed every Zande does most carefully, and that if he had not been bewitched he would have seen the stump. As a conclusive argument for his view he remarked that all cuts do not take days to heal but, on the contrary, close quickly, for this is the nature of cuts. Why, then, has his sore festered and remained open if there were no witchcraft behind it? (Evans-Pritchard 1976, p. 304)

In this passage, the boy reasons to the conclusion that witchcraft had made him knock his foot against a stump and prevented the wound from healing, and Pritchard records many structurally similar cases from his study of the Zande people. In each case, what comes to the fore is an event that deviates from the way things are supposed to go in some peculiar fashion that is to one's detriment. In the Zande culture, there is a general background belief that misfortune is often caused by witchcraft, and thus, it is easy to map the unseen forces of witchcraft onto the observable features of the misfortune.

Commenting on the case of the boy from the cited story, Evans-Pritchard says,

> The boy who knocked his foot against a stump of wood did not account for the stump by reference to witchcraft, nor did he suggest that whenever anybody kicks his foot against a stump it is necessarily due to witchcraft, nor yet again did he account for the cut by saying that it was caused by witchcraft, for he knew quite well that it was caused by the stump of wood. What he attributed to witchcraft was that on this particular occasion, when exercising his usual care, he struck his foot against a stump of wood, whereas on a hundred other occasions he did not do so, and that on this particular occasion the cut, which he expected to result from the knock, festered whereas he had dozens of cuts which had not festered. Surely these peculiar conditions demand an explanation (ibid, p. 305).

What is the boy doing? He isolates the peculiar features of the event, which include stubbing his toe while being attentive and the resultant cut being especially slow to heal. He then privileges these features to the exclusion of all the other details that surround the event. He, thereby, creates a model of what is salient at the level of the data to be explained. He then identifies a fit between his model of what has happened to him and the rather general, cultural model according to which bad things that happen, despite responsible behavior, are due to witchcraft. In the absence of a practical reason to pursue a deeper explanation, the boy is satisfied with the general match between the circumstances of his injury and the category of witchcraft.

The exact form that model-based religious reasoning takes will of course depend on all manner of circumstances, such as the kinds of extra-natural agents a given culture takes to exist, and under what conditions one can expect the interests of agents of those kinds to intersect with one's own interests. If someone believes in the prevalence of witchcraft, she might expect that any extraordinary event that would be in the interests of a normal human agent to bring about, but which seems impossible to account for in terms of a normal human agent exercising normal human powers, is a candidate for extra-natural explanation. By way of contrast, a consistent deist may make use of the model of a designer or creator in explaining some of the world's features, but her operative model of the world cannot permit the introduction of special divine action as a possible causal explanation of events in her environment on pain of inconsistency.

The form that model-based reasoning takes will also depend on what models are available as heuristics for applying the more complex, official doctrines of a religion, especially one for which the content of divine revelation cannot be captured adequately by any one model. For example, a Christian typically resorts to a wide range of metaphors and analogies for God, such as that of father, judge, creator, shepherd, wronged lover, warrior, and many other personal roles and even inanimate things, such as a rock or a consuming fire. Some of these models may be more appropriate than others in particular kinds of model-based reasoning, in particular circumstances. For example, the reasons that a judge would allow one to suffer, for example, could well differ from the reasons that a loving father would have for permitting suffering. Both these metaphors, as used in particular model-based reasoning processes, however, capture diverse facets of the officially sanctioned Christian understanding of the divine.

Nevertheless, although model-based reasoning in religious matters is extremely diverse, even within a single faith tradition, what does seem plausible is that some kind of modeling is required for any kind of explanation involving unseen intentional agency and, moreover, that the cognitive processes involved share at least some family resemblance with model-based reasoning in other areas of life, including the sciences. A religio-cultural superstructure may also be present and help to regulate the models of extra-natural agency people employ, as well as the degree of fit between such models and unusual events that is considered sufficient to justify acting on a perceived fit. There is reason to think that the basic models of folk psychology, however, may be so robust as to be very hard for a religious authority to constrain. For instance, religious people routinely reason with models of their gods that are anthropomorphic in respects not licensed by official doctrine (see, for example, Sloan (2004); Barrett and Keil (1996)).

Finally, the family resemblance of model-based reasoning in both religious and scientific reasoning also hints at some of the ways in which one might evaluate a specific instance of popular religious reasoning on its own terms. Some features of model-based reasoning that could be used to evaluate such reasoning as warranted or not include: whether one's degree of belief in the conclusion is proportional to the degree of fit between the model and the data; whether the former is sensitive to the latter or a match can be accounted for by accident; whether the pattern of one's

data is genuinely distinctive relative to competing models; whether one has been appropriately sensitive to other patterns that might be in the data; whether the model one applies to the data is internally consistent; whether the model is independently well-founded; and whether one has submitted one's own reasoning to whatever external checks may be available. For example, if we were to evaluate the reasoning of the Zande boy who cut his foot on the stump, we might have cause to criticize the boy's reasoning along one or more of these lines: we might point out that there does not seem to be any reason to expect that a witch would be interested in cutting the boy's foot, and thus, the fit between the anomalous experience and the model he is using may not be as tight as he thinks; alternatively, we might challenge how distinctive the boy's experience really is, or the general warrant for the witchcraft model of misfortune. What one should not do, however, is simply consign what the boy does to the category of "un-reason." Even if the boy's beliefs operate largely at a subliminal level, what the boy does is engage in a kind of reasoning, the pattern of which is not as unfamiliar as it might appear.

An Application to Natural Theology

As a coda to the previous sections, notice that model-based reasoning can work in either of two directions. The primary concern of a hiker is to find a map that she can trust and then to match her surroundings to the map so as to navigate the terrain. A mapmaker, on the other hand, works from the other direction, abstracting from his observations of the terrain to a composite representation of the whole.

As noted previously, the bidirectionality of model-based reasoning is present in the scientific and social domains as well as navigating physical maps. Model-based scientific reasoning need not be tied only to interpreting one's findings but can be used to generate and test new theories. James Snow's map of a cholera epidemic mentioned previously could have inspired a waterborne theory of cholera if Snow had not already formulated it. In the social domain, we do not simply interpret the behavior of persons in terms of our previous expectations. We construct new expectations on the basis of present behavior, and revise our models in light of the course of our relationship with others.

Natural theology, ultimately, tries to establish some theological truths with natural reason. In an early article in his *Summa theologiae* (*ST*), Aquinas takes up the question of whether it can be demonstrated that God exists. He replies in the affirmative:

> When an effect is better known to us than its cause from the effect we proceed to the knowledge of the cause. And from every effect the existence of its proper cause can be demonstrated, so long as its effects are better known to us; because since every effect depends upon its cause, if the effect exists, the cause must pre-exist. Hence the existence of God, in so far as it is not self-evident to us, can be demonstrated from those of his effects which are known to us (*ST* I q.2 a.2).

Aquinas goes on in the five ways to isolate different structural features of the natural world and to argue that they imply the existence of a certain kind of minimally described extra-natural being. Although Aquinas's famous proofs are deductive in form, in light of the discussion of model-based reasoning above, it is tempting to gloss what is happening here as "bottom-up" model-based reasoning. Just as Snow's map of the cholera outbreak identified a pattern in an effect that implied a corresponding structure in the cause of that effect ("something in the water causes cholera"), so Aquinas aims to demonstrate the existence of God under the minimal descriptions necessary for God to be the cause of a natural world with the structural properties Aquinas identifies: movement requires a first mover; an ordered series of causes needs an uncaused cause; teleology and purpose imply some undergirding intelligence that can imbue something with purpose and so on. For a rigorous reconstruction of Aquinas' five ways, see Pawl (2012).

Comparing this approach with the cases considered previously, my suggestion is that the way to see the contrast is that natural theology fits most appropriately with bottom-up model-based reasoning, whereas popular religious explanations are typically instances of top-down model-based reasoning. To express the contrast another way, natural theology is akin to map-making, whereas popular religion is akin to hiking with a preexisting map. Either can be done skillfully, but the former pairing is more specialized than the latter pairing. The average person has neither the time, the training, nor the motivation to attempt to generate or refine a model of the extra-natural. For ordinary interests, it is sufficient to take one's cultural inheritance and seek to apply it to one's circumstances to help navigate a path through life. This contrast, however, does not mean there is not some family resemblance between the two forms of religious reasoning or, indeed, between both of these forms and model-based reasoning in the sciences.

Conclusion and Future Directions

What model-based reasoning in the sciences, natural theology, and popular religious explanation all involve is navigation between the relational structures of the observable, natural world and whatever deep explanatory beings or structures are purported to lie beyond the limits of normal observation. Keeping this family resemblance in mind can help render intelligible, according to their own terms, processes of thought that many today might otherwise find especially peculiar about popular religious reasoning.

Some further questions that bear investigation are as follows.

Helen De Cruz has been doing groundbreaking work trying to give an explanation in terms of cognitive science of why certain traditional arguments for God's existence such as the teleological and cosmological arguments have perennial appeal (Cf. De Cruz 2010, 2011). Along these lines, there is a great deal of work to be done in analyzing the models and forms of reasoning that lie behind specific arguments and discussions within either popular religion or theology.

For instance, what are the models that might inform: a cyclical versus a linear view of time, rival perspectives on providence and free will, or various forms of theodicy?

Another question that merits investigation concerns the way in which one ends up deciding between rival models. The rival models could be found either between different religions, as when someone has a cultural background that allows them to slip between a cyclical and a linear view of time, or within a religion, as when one must decide to make sense of one's experience through a model of God as a loving father, a righteous judge, or a warrior battling evils.

Cross-References

► Cognition, Brain, and Religious Experience: A Critical Analysis
► Divine Understanding and the Divided Brain

References

Aquinas, T. (1911–1935). *The "Summa Theologica" of St. Thomas Aquinas, literally translated by the Fathers of the English Dominican Province*. London: Burns, Oates and Washbourne.

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion*. New York: Oxford University Press.

Barrett, J. (2004). *Why would anyone believe in God?* Lanham: Altamira Press.

Barrett, J., & Keil, F. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in god concepts. *Cognitive Psychology, 31*, 219–247.

Bering, J. (2011). *The god instinct: The psychology of souls, destiny, and the meaning of life*. New York: Norton.

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York: Basic Books.

De Cruz, H. (2010). Paley's ipod: The cognitive basis of the design argument within natural theology. *Zygon, 45*(3), 665–684.

De Cruz, H. (2011). The cognitive appeal of the cosmological argument. *Method & Theory in the Study of Religion, 23*(2), 103–122.

Evans-Pritchard, E. E. (1976). *Witchcraft, oracles, and magic among the Azande*. Oxford: Clarendon Press.

Frigg, R. & Hartmann, S. (2006/2012). Models in science. Stanford encyclopedia of philosophy. http://stanford.library.usyd.edu.au/entries/models-science. Accessed 15 Nov 2012.

Giere, R. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Giere, R. (2006). *Scientific perspectivalism*. Chicago: University of Chicago Press.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. New York: Oxford University Press.

Godfrey-Smith, P. (2005). Folk psychology as a model. *Philosophers Imprint, 5*, 1–16.

Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review, 112*, 468–493.

Guthrie, S. (1993). *Faces in the clouds: A new theory of religion*. New York: Oxford University Press.

Hodges, W. (1997). *A shorter model theory*. New York: Cambridge University Press.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Johnson, S. (2006). *The ghost map: The story of London's most terrifying epidemic—and how it changed science, cities, and the modern world*. New York: Riverhead Books.

Knauff, M., & May, E. (2006). Mental imagery, reasoning, and blindness. *The Quarterly Journal of Experimental Psychology, 59*, 161–177.

Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience, 15*, 559–573.

Maibom, H. (2009). In defence of (model) theory theory. *Journal of Consciousness Studies, 16*, 360–378.

McCauley, R. (2011). *Why religion is natural and science is not*. New York: Oxford University Press.

McCauley, R., & Lawson, E. T. (2002). *Bringing rituals to mind: Psychological foundations of cultural forms*. New York: Cambridge University Press.

Pawl, T. (2012). The five ways. In B. Davies & E. Stump (Eds.), *The Oxford handbook of Thomas Aquinas* (pp. 115–131). New York: Oxford University Press.

Simon, H. (1977). *Models of discovery and other topics in the methods of science*. Boston: D. Reidel.

Sloan, D. J. (2004). *Theological incorrectness: Why religious people believe what they shouldn't*. New York: Oxford University Press.

Thagard, P. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Divine Understanding and the Divided Brain

100

Iain McGilchrist

Contents

Abstract

Interaction with the world requires the right hemisphere's broad attention, which is inclusive and opens up into possibility, coupled with the left hemisphere's narrow attention, which collapses the world we experience into specificity. If the left hemisphere collapses the world too quickly into what is specific, however, it precludes the possibility of knowledge that transcends what is already familiar, notably purported knowledge of the divine. By contrast, the right hemisphere is more sensitive to image, metaphor, and narratives by which theological knowledge may be capable of expression that would be ambiguous or apparently contradictory if expressed simply as a set of propositions. The reciprocal organization of the cerebral hemispheres therefore suggests that any proposed theology that is articulated simply in terms of a set of specific propositions about the divine risks betraying, distorting, and misrepresenting its subject matter. Furthermore, this organization of brain structure and function suggests that images, metaphors, and narratives are not poor substitutes or intermediate steps in theological knowledge, but indispensable to it.

I. McGilchrist
The Bethlem Royal and Maudsley Hospital, London, UK
e-mail: iain.mcgilchrist@all-souls.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_99,

Introduction

How does one know, or at any rate become aware of, what it is that one does not know? This is, of course, a fundamental epistemological problem, since the openness required to embrace propositions that cannot be demonstrated to follow necessarily from commonly held truths leaves one open to the possibility of self-deceit. And yet not to be aware of what it is one does not know is also, inevitably, to deceive oneself.

Language and Hemisphere Differences

Not all knowledge is, however, propositional in nature. Ultimately, in fact, all knowledge derives from experience, for which there are no propositions, and many languages express at least some aspect of this difference by distinct words for the ways in which one can be said "to know." *Sapere* bows to *cognoscere*, *savoir* to *connaître*, *wissen* to *kennen*. When we say we know something, what we mean is that we see that it is like something else that we reckon we already know better. And those "somethings else," followed far enough, return us in every case to embodied experience. Additionally, both the first propositions, or axioms, from which we reason, and even the value of reason itself as a tool for the discovery of truth, have to be intuited. We cannot reason our way to either.

Since language embodies thought, it follows the same path. Words are like money. In any (apparently) enclosed financial system, any number of "virtual" transactions can be carried out, but in the end, all such transactions depend on money taking its value from somebody's cows or chickens somewhere, and being translated back into real goods or services – food, clothes, car repairs – in the realm of daily life somewhere else. So it is with words. Webs of thought can be spun with them, but ultimately language represents something valuable elsewhere: Its value lies only in what it represents in the world of embodied experience.

For this reason, language is essentially, not accidentally, metaphorical in nature. Even the most abstract of terms, such as the word "virtual" itself, take us back ultimately to the earthy reality of a man's strength (*vir-tus* in Latin). Metaphor embodies thought and places it, where it belongs, in a living context. In this, it bridges the gap between language and the world, a gap entailed on us by the very nature of language. The language of science and philosophy is not only no exception to this, but a particularly obvious example of it.

Thus both the process of reason and its axioms, and the business of linguistic discourse and its terms, ultimately depend on and cannot transcend intuitive knowledge and embodied experience. Following intuition may lead us astray, right enough, but so may not attending to it: There is a wealth of evidence that those who rely on ratiocination alone make poorer judgments than those who combine reason with intuition, and often they can barely function in the world at all.

What does all this have to do with the brain? In humans, as in other mammals, as well as in birds, reptiles, and even fish, the brain is divided. This is odd, because the brain exists to make connections and is only as powerful as its connections. Odder still, the band of fibers that connects the hemispheres, the corpus callosum, has grown proportionally smaller (in relation to hemispheric volume) with evolution (Jäncke and Steinmetz 2003; Hopkins and Marino 2000; Aboitiz et al. 1992), and much of its activity in any case involves functional inhibition of the contralateral hemisphere. The corpus callosum contains an estimated 300–800 million fibers connecting topologically similar areas in either hemisphere. Yet only 2 % of cortical neurons are connected by this tract (Jäncke and Steinmetz 2003; Banich 2003). What is more, the main purpose of a large number of these connections is actually to inhibit – in other words to stop the other hemisphere interfering. Neurons can have an excitatory or inhibitory action, excitatory neurons causing further neuronal activity downstream, while inhibitory neurons suppress it. Although the majority of cells projecting to the corpus callosum use the facilitatory neurotransmitter glutamate, and are excitatory, there are significant populations of nerve cells (those that use the neurotransmitter gamma-amino butyric acid, GABA) whose function is inhibitory. Even the excitatory fibers often terminate on intermediary neurons, or interneurons, whose function is inhibitory (Conti and Manzoni 1994; Saron et al. 2002). Inhibition is, of course, not a straightforward concept. Inhibition at the neurophysiological level does not necessarily equate with inhibition at the functional level, any more than letting your foot off the brake pedal causes the car to halt: Neural inhibition may set in train a sequence of activity, so that the net result is functionally permissive. But the evidence is that the primary effect of callosal transmission is to produce *functional* inhibition (Meyer et al. 1995; Röricht et al. 1997; Höppner et al. 1999). So much is this the case that a number of neuroscientists have proposed that the whole purpose of the corpus callosum is to allow one hemisphere to inhibit the other (Cook 1984; Hoptman and Davidson 1994; Chiarello and Maxfield 1996). Stimulation of neurons in one hemisphere commonly results in an initial brief excitatory response, followed by a prolonged inhibitory arousal in the contralateral hemisphere. Such inhibition can be widespread, and can be seen on imaging (Saron et al. 2003; Allison et al. 2000; Tootell et al. 1998).

There would appear to be an evolutionary adaptation here that connects, but also importantly separates, two spheres of cerebral activity. Why?

Birds and other animals have to solve a conundrum on which their survival depends, namely, how to eat and to stay alive at the same time. Each must pay attention to something that is already prioritized – a seed, one's prey – at the same time as being open to whatever it is that might come along during the process – be it predator or conspecific. For the first of these, one needs narrow-beam, sharply focused attention to something that is already prioritized; for the latter, one needs precisely the opposite, namely, a broad, open, vigilant, sustained attention without commitment as to what may be found. Paying two kinds of attention in one consciousness at the same time is an almost intractable problem. The solution appears to have been the bihemispheric brain. The left hemisphere in birds and other animals provides narrow attention in order to get food, to pick up a twig to

build a nest, and in general to *manipulate* the world; the right hemisphere provides a broad picture that makes it possible to watch out for predators and bond with mates, and more generally to *understand*, and to find oneself standing in relation to, the world at large. Unsurprisingly, therefore, chicks that are properly lateralized (whose hemispheres are appropriately differentiated) are more able to use these two types of attention effectively than are those in whom, experimentally, lateralization has not been permitted to develop (Rogers 2000). Many types of bird show more alarm behavior when viewing a predator with the left eye (right hemisphere) (Hoffman et al. 2006), are better at detecting predators with the left eye (Rogers and Kaplan 2006; Rogers 2000), and will choose to examine predators with their left eye (Rogers et al. 2004), to the extent that if they have detected a predator with their right eye, they will actually turn their head so as to examine it further with the left (Dharmaretnam and Rogers 2005). Hand-raised ravens will even follow the direction of gaze of a human experimenter looking upward, using their left eye (Bugnyar et al. 2004). For many animals, there are biases at the population level toward, again, watching out for predators with the left eye (Evans et al. 1993; Rogers 2000; Lippolis et al. 2002, 2005). In marmosets, individual animals with more strongly lateralized brains are better able, because of hemisphere specialization, to forage and remain aware of predators (Rogers 2005). There are shorter reaction times in cats that have a lateralized paw preference (Fabre-Thorpe et al. 1993). Lateralized chimps are more efficient at fishing for termites than unlateralized chimps (McGrew and Marchant 1999). Even individual human brains that are, for one reason or another, less lateralized than the norm appear to show global deficits (Crow et al. 1998). In a word, lateralization brings evolutionary advantages, particularly in carrying out dual-attention tasks (Rogers et al. 2004). As one researcher has put it succinctly: Asymmetry pays (Güntürkün et al. 2000).

In predatory birds and animals, it is the left hemisphere that latches on, through the right eye and the right foot, to the prey (Csermely 2004). It is certainly true of familiar prey: In toads, a novel or unusual choice of prey may activate the right hemisphere, until it becomes familiar as an object of prey, when it once again activates the left (Robins and Rogers 2006). In general, toads attend to their prey with the left hemisphere, but interact with their fellow toads using the right hemisphere (Vallortigara et al. 1998).

The advantages accrue not only to the individual: Being a more lateralized species at the population level carries advantages in social cohesion (Bisazza et al. 2000; Rogers and Workman 1989; Halpern et al. 2005). That may be because the right hemisphere appears to be deeply involved in social functioning, not just in primates, where it is specialized in the expression of social feelings, but in lower animals and birds as well (Fernández-Carriba et al. 2002; Ventolini et al. 2005). For example, chicks preferentially use the left eye (right hemisphere) for differentiating familiar members of the species from one another, and from those who are not familiar, and in general for gathering social information (Rogers 2000; Vallortigara 1992). Chicks approach their parents or an object on which they have imprinted using their left eye (Dharmaretnam and Andrew 1994), as do Australian magpies (Hoffman et al. 2006). Though black-winged stilts peck more, and more

successfully, at prey using the right eye (left hemisphere), males are more likely to direct courtship displays to females that are seen with their left eye (right hemisphere) (Ventolini et al. 2005). The right hemisphere is the main locus of early social experience in rats (Denenberg et al. 1978). In most animal species, intense emotional responses are related to the right hemisphere and inhibited by the left (Andrew and Rogers 2002).

In humans, too, the hemispheres attend to the world differently. Since attention is involved in the genesis of our experiential world, so that the quality of attention we pay affects what it is that we find, this has important consequences for the type of world each hemisphere helps to mediate for us.

Attention is not just another "function" alongside other cognitive functions. Its ontological status is of something prior to functions and even to things. The kind of attention we bring to bear on the world changes the nature of the world we attend to, the very nature of the world in which those "functions" would be carried out, and in which those "things" would exist. Attention changes *what kind of* a thing comes into being for us: In that way, it changes the world. If you are my friend, the way in which I attend to you will be different from the way in which I would attend to you if you were my employer, my patient, the suspect in a crime I am investigating, my lover, my aunt, a body waiting to be dissected. In all these circumstances, except the last, you will also have a quite different experience not just of me, but of yourself: You would feel changed if I changed the type of my attention. And yet nothing *objectively* has changed.

So it is, not just with the human world, but with everything with which we come into contact. A mountain, that is a landmark to a navigator, a source of wealth to the prospector, a many-textured form to a painter, or to another the dwelling place of the gods, is changed by the attention given to it. There is no "real" mountain which can be distinguished from these, no one way of thinking which reveals the true mountain.

It is often wrongly thought, however, that science uncovers such a reality. Its apparently value-free descriptions are assumed to deliver *the* truth about the object, onto which our feelings and desires are later painted. Yet this highly objective stance, this "view from nowhere," to use Nagel's phrase, is itself value-laden (Nagel 1986). It is just one particular way of looking at things, a way which privileges detachment, a lack of commitment of the viewer to the object viewed. For some purposes, this can be undeniably useful. But its use in such causes does not make it truer or more real, closer to the nature of things.

Attention also changes who *we* are, we who are doing the attending. Our knowledge of mirror neurons (Rizzolatti et al. 2001) and their function and of the effects of association-priming (Dijksterhuis et al. 2000) shows that by attending to someone else performing an action, and even by thinking about them doing so – even, in fact, by thinking about certain sorts of people at all – we become objectively, measurably, more *like* them, in how we behave, think, and feel. Through the direction and nature of our attention, we prove ourselves to be partners in creation, both of the world and of ourselves. In keeping with this, attention is inescapably bound up with value – unlike what we conceive as "cognitive

functions," which are neutral in this respect. Values enter only through *the way in which* such functions are exercised: They can be used in different ways for different purposes to different ends. Attention, however, intrinsically is a *way in which*, not a thing: It is intrinsically a relationship, not a brute fact. It is a "howness," a something between, an aspect of consciousness itself, not a "whatness," a thing in itself, an object of consciousness. It brings into being a world and, with it, depending on its nature, a set of values.

In experience, these two versions of the world, that of the right hemisphere and of the left, are merged or rapidly alternated in such a way that the subject is not aware of the fact, hence the significance of what we can learn from studying the brain hemispheres: It draws attention to an aspect of the structure of the phenomenal world that otherwise might elude us. The differences that emerge between the two hemispheres in their mode of attending to the world help us to understand better not only the experiential world, but understanding itself.

Some important caveats should be entered at this point. First, it should be clear that these differences are not absolute, but relative, and there is overlap between hemispheres in most respects. Nonetheless, the differences remain consistent and significant. Innate differences are subsequently amplified through experience, since expediency dictates that even a small advantage for one hemisphere in dealing with a certain kind of experience results in its being preferentially used to deal with similar experiences in future. In this way, the hemispheres become further differentiated during development. Second, although general truths are approximate and cannot be taken for rules, they are necessary for understanding. Contrary findings will inevitably exist to any such generalities. However, general truths are no less important for that. The average temperatures in Iceland and Indonesia are clearly very different, which goes a long way to explain the wholly different characteristics of the vegetation, animal life, landscape, culture, and economy of these two regions, as well as no doubt much else that differentiates their "feel" and the ways of life there. But it is still true that the lowest average annual temperature in Indonesia is *lower* than the highest average annual temperature in Iceland – and of course the average temperature varies considerably from month to month, as well as, less predictably, from day to day, and indeed from place to place within each region. This leads to a third caveat: There are interindividual differences, and individuals vary from occasion to occasion.

It is also true that a left hemisphere or right hemisphere on its own cannot be said to do what only a person can do: "believe," "intend," "decide," "prefer," and so on. These and similar formulations should be understood as avoiding the repetition of such cumbersome locutions as "a subject relying on the cognitive faculties of the left [or right] hemisphere believes," etc.

For a full account of the nature of the differences between the hemispheres, the reader is referred to *The Master and his Emissary*: *The Divided Brain and the Making of the Western World* (McGilchrist 2009). However some broad, "headline" distinctions could be made here. It should be understood, nonetheless, that these are very general distinctions, distinctions that ask for qualification at a length that is prohibited here.

The left hemisphere's world requires precision rather than breadth, and aims to close things down as much as possible to a certainty, where the right hemisphere views the broad picture and opens things up to possibility (Ivry and Robertson 1998; Kitterle et al. 1990; Sergent 1982; Robertson et al. 1988; Robertson and Lamb 1991; van Kleeck 1989). The right hemisphere is more capable of a frame shift than the left (Rausch 1977). In focusing on its object, the left hemisphere renders it explicit, and abstracts it from its context (Kinsbourne 1988; Federmeier and Kutas 1999): The right hemisphere is aware of, and able to deal appropriately with, all those things that are required to remain implicit, and are denatured once removed from their context (Alexander et al. 1989; Heilman et al. 1975). The left hemisphere conceives of its object as static, fixed, and atomistic, rather than, as the right hemisphere does, fluid, evolving and interconnected with the rest of the world (Cummings 1997; Bender et al. 1968; Michel and Troost 1980; Müller et al. 1995; Corballis 1996; Corballis et al. 1998). Where the left hemisphere sees disconnected fragments from which the whole picture might be constructed, the right hemisphere sees the whole, the Gestalt, which is more than the sum of the parts, and from which the "parts" have artificially to be determined (Yoshida et al. 2007; Evert and Kmen 2003; Fink et al. 1999; van Kleeck 1989). The right hemisphere is alive to what "presences" to us (to use a Heideggerian term) pre-conceptually; the left hemisphere deals with what is already familiar as a "re-presentation," literally present after the fact (Goldberg 1990; Goldberg and Costa 1981). Where the left hemisphere sees things as general, and disembodied or abstract, the right hemisphere sees them as unique, incarnate, and concrete (Cutting 1997; Kosslyn 1987; Goldberg 1990; Hécaen and Albert 1978). If the left hemisphere is concerned with what can be counted, the quantitative and measurable aspect of experience, the right hemisphere is concerned with the qualitative (Marsolek 1995; Brown and Kosslyn 1993; Kosslyn 1987; Grossman 1988; Cutting 1997; Warrington and McCarthy 1987; Gardner 1974; Bornstein et al. 1969; Bornstein and Kidron 1959; Landis et al. 1986; Bourgeois et al. 1998). One could say that the right hemisphere's world is living, whereas the left hemisphere's world is mechanical and inanimate (Corballis 1998) – for example, only the left hemisphere codes for tools and machines (Gainotti 2002; Perani et al. 1995; Martin et al. 1996; Price and Friston 2002; Mummery et al. 1996; Damasio et al 1996; Cutting 1997).

One might expect, on the basis of the above, that the ways in which the two hemispheres communicated their experience of the world would also differ. It is well known that the left hemisphere alone, in the vast majority of right-handers, is capable of speech in language, but less well known that both hemispheres are intimately involved with language reception – the right hemisphere being especially important for the understanding of an utterance as a whole, in context, with all its nonliteral, implicit meaning (Foldi et al. 1983; Kaplan et al. 1990; Heilman et al. 1975). The right hemisphere therefore plays an important part in the understanding of linguistic utterance, but does not favor language as a way of expressing its perceptions of the world.

By contrast, music is the natural form of expression in the right hemisphere, and the musical aspects of language, including pitch excursions, intonation, inflection,

and rhythm, along with facial expression and body language, all of which are underwritten by the right hemisphere (Blonder et al. 1991; Borod 1993; Breitenstein et al. 1998; Ross et al. 1977; Haggard and Parkinson 1971; Carmon and Nachshon 1973; Wymer et al. 2002; Cutting 1997; Blakeslee 1980), are particularly important in the second-person, "I-Thou" relationship, though they become less important in the third-person, "I-It" relationship. Many contemporary anthropologists indeed argue that music was the primary mode of human expression, and that language emerged from music relatively late in human social evolution, between 80,000 and 40,000 years BC, possibly as a response to the need to communicate explicitly once societies grow beyond a certain point, and have instrumental needs that require referring to things, places, and people not present to the speakers (Milo and Quiatt 1993; Mithen 1998, 2005).

One of the most significant differences between the approaches of the two hemispheres to language is that it is the right hemisphere that is best able to understand metaphor (Foldi 1987; Bottini et al. 1994). There is however some confusion over what is meant by metaphor. Obviously, there is metaphoric content to almost everything we say – language is essentially metaphoric in nature, at the simplest level. In an extensive literature which confirms the right hemisphere's key role in understanding metaphor, there are two studies (Rapp et al. 2004; Stringaris et al. 2007) which have suggested that it is the left hemisphere that is principally involved in the appreciation of metaphor. However, these studies used only overfamiliar or hackneyed expressions. When the metaphor is new or imaginatively demanding, the kind encountered in poetry rather than cliché, it is clearly the right hemisphere that is involved (Faust and Mashal 2007; Foldi et al. 1983; Kaplan et al. 1990). Thus, poetic phrases, such as "rain clouds are pregnant ghosts," are understood by the right hemisphere, while clichés, such as, "babies are angels," are understood by the left hemisphere (Schmidt et al. 2007). Familiar expressions activate the left hemisphere, whereas unfamiliar ones activate the right hemisphere (Bottini et al. 1994; Eviatar and Just 2006; Mashal et al. 2005, 2007). However, it is not the novelty effect alone, but specifically the combination of novelty with metaphorical content that involves the right hemisphere (Mashal and Faust 2008).

Metaphor reveals connections: It relies on openness to the fruitfulness of ambiguity, and the recognition of knowledge which is not propositional in nature but based on relations between forms. In fact, all the aspects of language that are peculiarly important to poetry – not just metaphor, but implied meaning of every kind, irony, connections between ideas not normally approximated, the connotative power of symbols, the music of language (the movement of verse, its ictus, meter, and rhyme) – are all preferentially mediated by the right hemisphere. By common consent over generations, it is poetry which enshrines our profoundest insights into reality, and even philosophers such as Wittgenstein and Heidegger came to believe that philosophical discourse needed ultimately to cede to poetry, as Schopenhauer believed it needed to cede to music. Propositional discourse is limited in its ability to approach ultimate reality. Its very terms take us back, as left hemisphere discourse always does, to the familiar. In Nietzsche's words, language makes the uncommon common.

Additionally, and as might be expected from all the above, the right hemisphere is better able than the left to understand the meaning of symbols with implicit, multiple meanings (for example, the rose, as opposed to the explicit meaning or 1:1 mapping of "red" onto "stop" in the case of a traffic light) (Gloning et al. 1968; Goldberg 1990). In other words, it understands the complex of connotations and how they work to enrich meaning, whether this is in words, or images, or in symbolic enactment, such as ritual of all kinds. The left hemisphere sees in such phenomena only a lack of precision: It sees obfuscation or, at worst, untruth, hence the Enlightenment view expressed by Locke that metaphors were "perfect cheats" (Locke 1690). The right hemisphere is also more able to understand narrative (Ornstein et al. 1979; Mills and Rollman 1980; Swisher and Hirsh 1972; Carmon and Nachshon 1971; Nicholls 1994; Brown and Nicholls 1997; Hough 1990; Schneiderman et al. 1992; Vogeley et al. 2001): The unaided left hemisphere tends to categorize similar episodes of a story together and get them out of sequence, because it does not follow the overall meaning, the way in which a human story unfolds (McNeill 1992).

What does this tell us when we return to the question of how we may avoid the complacent belief that we know all there is to know? Every indicator is that the right hemisphere both grounds our experience of the world, at the bottom end, so to speak, and makes sense of it, at the top end. Broad vigilant attention is primary: Though focused attention may appear to its owner to be under conscious control, in reality, it is already spoken for, since we direct attention according to what we are aware of, and for that we need broad, right hemisphere, attention. Then there is the primacy of wholeness: The right hemisphere deals with the world before separation, division, and analysis have transformed it into something else, before the left hemisphere has re-presented it. It is not that the right hemisphere connects – because what it reveals was never separated; it does not synthesize – what was never broken down into parts; it does not integrate what was never less than whole. The right hemisphere also delivers what is new: It has the primacy of experience. And we must confront the fact that the left hemisphere's most powerful tool, referential language, has its origins in the body and the right hemisphere – a sort of primacy of means. Equally the implicit comes before the explicit. The right hemisphere is more in touch with both affect and the unconscious will, and neurological evidence supports what is called the primacy of affect and the primacy of unconscious over conscious will. Most remarkable of all, some subtle work by David McNeill shows that thought originates in the right hemisphere and is only "worked up" by the left hemisphere (McNeill 1992). For understanding, what is then produced by the left hemisphere needs to be returned to the whole context known only to the right.

The left hemisphere may add – and it adds enormously much – but a return is required to the world that is grounded by the right hemisphere. The left hemisphere's world is a virtual world, one where we are no longer patient recipients, but powerful operators. The values of clarity and fixity are added through the processing carried out by the left hemisphere, making it possible for us to control, manipulate, and use the world. To this end, attention is directed and focused; the

wholeness is broken into parts; the implicit is unpacked; language becomes the instrument of serial analysis and things are categorized and become familiar. In addition, affect is set aside and superseded by cognitive abstraction; the conscious mind is brought to bear on the situation; thoughts are sent to the left hemisphere for expression in words and the metaphors are temporarily lost or suspended, the world being re-presented in a static and hierarchically organized form. This re-presentation enables us to have knowledge, to bring the world into resolution, but it leaves what it knows denatured and decontextualized. For understanding rather than efficient manipulation, there needs to be a process of reintegration whereby we return to the experiential world again. The parts, once seen, are subsumed again in the whole, as the musician's painful, conscious, fragmentation of the piece in practice is lost in the (now improved) performance. The part that has been under the spotlight is seen as belonging to a broader picture; what had to be conscious for a while becomes unconscious again; what needs to be implicit once again retires; the represented entity becomes once more present, and "lives"; and even language is given its final meaning by the right hemisphere's holistic pragmatics.

All the evidence is that the right hemisphere "sees" more than the left hemisphere. Perhaps it is for this very reason that it is more aware of the limitations of its knowledge. In neuropsychological studies, the right hemisphere exhibits a more tentative and self-depreciating style, whereas the left hemisphere is confident about matters of which it is ignorant, and overestimates its capacities (Gazzaniga and LeDoux 1978; Cutting 1997; Rausch 1985; Brownell et al. 1986; Henson et al. 2000; Kimura 1963; Phelps and Gazzaniga 1992; Schnider et al. 1996; Nebes 1974; Drake and Bingham 1985). Moreover, the left hemisphere is better able to inhibit the right than the right hemisphere is able to inhibit the left (Kinsbourne 1993; Oliveri et al. 1999), and transmission is faster from right hemisphere to left than from left hemisphere to right (Marzi et al. 1991; Bisiacchi et al. 1994; Brown et al. 1994; Saron and Davidson 1989). It seems that the cognitive processes of the right hemisphere allow it to appreciate the importance of *what it does not know*, whereas those of the left hemisphere do not permit such insight. And indeed, insight into one's capabilities is largely dependent on the right frontal region (Adair et al. 2003; Bisiach et al. 1986; Feinberg et al. 1994; Meador et al. 2000; Starkstein et al. 1992; Stuss 1991).

Whatever we may mean by the realm of the divine, any apprehension of it would be rendered improbably difficult by a cognitive-perceptual style, the aim of which was to narrow things down to a certainty, and to "recognize" them as examples of something already familiar. More particularly, such a realm would be recalcitrant to thought processes that require explicitness and see truth as a property of a set of syllogisms that exhibit internal coherence. Many – perhaps all – aspects of the human world that give meaning to life present such difficulties. To take just one example, the experience of love in all its many kinds is such that, despite its supreme importance, it cannot be well conveyed in language, especially not in propositional language. Such experiences are part of our embodied life taken as a whole. They can be conveyed best by poetry, or drama, or ritual, or image, or narrative, or music – by means, in other words, that are implicit, embodied, and

contextually rich. They are resonant rather than declarative. The inherent weakness of the analytic method applied to theological matters has been described as "cognitive *hemianopia*" (Stump 2009). This nicely hints that only half the visual field, and therefore one hemisphere's view, is being taken into account. It is not that we are using the proper faculties for the task, though they are half-blind, but that we are using the wrong faculties altogether. Trying to approach the divine using the serial, analytic methods of the left hemisphere is like trying to discover whether the sun is shining by listening for the noise it makes.

At the center of the problem is that although it may be convenient, for some purposes, to pretend that some things are inert, distinct, and atomistic, everything actually exists in a network of relationships with everything else. Some things, such as music, exist only in relationships and yet have the power to mean as much as anything in the world. Music consists entirely of relations, "betweenness." The notes mean nothing in themselves: The tensions between the notes, and between notes and the silence with which they live in reciprocal indebtedness, are everything. Melody, harmony, and rhythm each lie in the gaps, and yet, the betweenness is only what it is because of the notes themselves. Actually the music is not *just* in the gaps any more than it is *just* in the notes: It is in the whole that the notes and the silence make together. Each note becomes transformed by the context in which it lies. What we mean by music is not just any agglomeration of notes, but one in which the whole that is created is powerful enough to make each note live in a new way, a way that it had never done prior to this creation. Similarly poetry cannot be just any arrangement of words, but one in which each word is taken up into the new whole and made to live again in a new way, carrying us back to the world of experience, to life: Poetry constitutes a "speaking silence." Music and poetic language are both part of the world that is delivered by the right hemisphere, the world characterized by betweenness. Perhaps it is not, after all, so wide of the mark to call the right hemisphere the "silent" hemisphere: Its utterances are implicit.

Similarly we know that, while Newtonian mechanics operates at a purely local level, the physical universe is not precise, but probabilistic, that particles separated by a universe may exhibit entanglement, and that the act of observation affects the observed. Although direct comparisons between physics and psychology might appear somewhat far-fetched, the point is a more limited one: Simply that the precision, isolation, stasis, linearity, and predictability of the Newtonian universe is a rough and ready approximation, which works well enough for practical purposes, but is not actually true to the nature of reality, even of all physical reality. Against a space-time background, we model the world as simple two-body systems close to equilibrium, systems which are idealized and only imperfectly representative of the real world: a model that the left hemisphere constructs to enable us to manipulate the world. This representation is opposed to the right hemispheric apprehension of a world that is living, constantly flowing and changing, radically interconnected within itself, and unpredictable – a world that is truer not only to what contemporary physics tells us, but to what attentive reflection on the nature of the experienced world would tell us, were we not conditioned to think of physical reality in terms derived from a model of the universe discredited over a century ago.

In sum, we will never achieve even the starting point for an understanding of the divine by approaching it with the tools of the left hemisphere. Such tools have evolved for purely practical purposes – those of using and manipulating the world expediently. Instead, I suggest we need to approach purported revelations or self-disclosures of the divine principally with the tools of the right hemisphere, the one whose purpose is to be on the lookout for what is not suspected by the left hemisphere, busily engaged as this hemisphere is in the local matter of getting and using. The right hemisphere has evolved to help us be aware of what is going on in the world as a whole, to understand the nature of things in the context of the whole, and to guide our relationship with them. This means that any proposed theology, that is articulated simply in terms of a set of specific propositions drawn from a self-enclosed, abstract model of the world, prizing clarity and internal consistency above all else, risks betraying, distorting, and misrepresenting its subject matter. As the apophatic tradition suggests, such propositions will be inevitably untrue, and stand between us and a purer understanding.

Conclusion and Future Directions

Insights from contemporary neuroscience confirm that images, metaphors, and narratives are indispensable to theological knowledge, not just poor substitutes or intermediate steps in its evolution. A worldview that places supreme value on what is certain, explicit, and literal is likely to misunderstand the nature of theology. At the same time, it will be unaware of what it is that it does not and cannot see. A deeper understanding of the differentiation of the cerebral hemispheres may contribute to debates on the nature of understanding itself.

Cross-References

- ▶ Cognition, Brain, and Religious Experience: A Critical Analysis
- ▶ Explanation and Levels in Cognitive Neuroscience
- ▶ The Contribution of Neurological Disorders to an Understanding of Religious Experiences
- ▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Aboitiz, F., Scheibel, A. B., & Zaidel, E. (1992). Morphometry of the Sylvian fissure and the corpus callosum, with emphasis on sex differences. *Brain, 115*(5), 1521–1541.

Adair, J. C., Schwartz, R. L., & Barrett, A. M. (2003). Anosognosia. In K. M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (4th ed., pp. 185–214). Oxford: Oxford University Press.

Alexander, M. P., Benson, D. F., & Stuss, D. T. (1989). Frontal lobes and language. *Brain and Language, 37*(4), 656–691.

Allison, J. D., Meador, K. J., Loring, D. W., et al. (2000). Functional MRI cerebral activation and deactivation during finger movement. *Neurology, 54*(1), 135–42.

Andrew, R. J., & Rogers, L. J. (2002). The nature of lateralisation in tetrapods. In R. J. Andrew & L. J. Rogers (Eds.), *Comparative vertebrate lateralisation* (pp. 94–125). Cambridge: Cambridge University Press.

Banich, M. T. (2003). Interaction between the hemispheres and its implications for the processing capacity of the brain. In K. Hugdahl & R. J. Davidson (Eds.), *The asymmetrical brain* (pp. 261–302). Cambridge, MA: Massachusetts Institute of Technology Press.

Bender, M. B., Feldman, M., & Sobin, A. J. (1968). Palinopsia. *Brain, 9*(2), 321–338.

Bisazza, A., Cantalupo, C., Capocchiano, M., et al. (2000). Population lateralisation and social behaviour: A study with sixteen species of fish. *Laterality, 5*(3), 269–284.

Bisiacchi, P., Marzi, C. A., & Nicoletti, R. (1994). Left-right asymmetry of callosal transfer in normal human subjects. *Behavioural Brain Research, 64*(1–2), 173–178.

Bisiach, E., Vallar, G., Perani, D., et al. (1986). Unawareness of disease following lesions of the right hemisphere: Anosognosia for hemiplegia and anosognosia for hemianopia. *Neuropsychologia, 24*(4), 471–482.

Blakeslee, T. R. (1980). *The right brain*. London: Macmillan.

Blonder, L. X., Bowers, D., & Heilman, K. M. (1991). The role of the right hemisphere in emotional communication. *Brain, 114*(3), 1115–1127.

Bornstein, B., & Kidron, D. P. (1959). Prosopagnosia. *Journal of Neurology, Neurosurgery, and Psychiatry, 22*(2), 124–131.

Bornstein, B., Sroka, H., & Munitz, H. (1969). Prosopagnosia with animal face agnosia. *Cortex, 5*(2), 164–169.

Borod, J. C. (1993). Cerebral mechanisms underlying facial, prosodic, and lexical emotional expression: A review of neuropsychological studies and methodological issues. *Neuropsychology, 7*, 445–463.

Bottini, G., Corcoran, R., Sterzi, R., et al. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language: A positron emission tomography activation study. *Brain, 117*(6), 1241–1253.

Bourgeois, M. J., Christman, S., & Horowitz, I. A. (1998). The role of hemispheric activation in person perception: Evidence for an attentional focus model. *Brain and Cognition, 38*(2), 202–219.

Breitenstein, C., Daum, I., & Ackermann, H. (1998). Emotional processing following cortical and subcortical brain damage: Contribution of the fronto-striatal circuitry. *Behavioural Neurology, 11*(1), 29–42.

Brown, H. D., & Kosslyn, S. M. (1993). Cerebral lateralisation. *Current Opinion in Neurobiology, 3*(2), 183–186.

Brown, S., & Nicholls, M. E. (1997). Hemispheric asymmetries for the temporal resolution of brief auditory stimuli. *Perception and Psychophysics, 59*(3), 442–447.

Brown, W. S., Larson, E. B., & Jeeves, M. (1994). Directional asymmetries in interhemispheric transmission time: Evidence from visual evoked potentials. *Neuropsychologia, 32*(4), 439–448.

Brownell, H. H., Potter, H. H., Bihrle, A. M., et al. (1986). Inference deficits in right brain-damaged patients. *Brain and Language, 27*(2), 310–321.

Bugnyar, T., Stöwe, M., & Heinrich, B. (2004). Ravens, Corvus corax, follow gaze direction of humans around obstacles. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 271*(1546), 1331–1336.

Carmon, A., & Nachshon, I. (1971). Effects of unilateral brain damage on the perception of temporal order. *Cortex, 7*(4), 411–418.

Carmon, A., & Nachshon, I. (1973). Ear asymmetry in perception of emotional nonverbal stimuli. *Acta Psychologica, 37*(6), 351–357.

Chiarello, C., & Maxfield, L. (1996). Varieties of interhemispheric inhibition, or how to keep a good hemisphere down. *Brain and Cognition, 30*(1), 81–108.

Conti, F., & Manzoni, T. (1994). The neurotransmitters and postsynaptic actions of callosally projecting neurons. *Behavioural Brain Research, 64*(1–2), 37–53.

Cook, N. D. (1984). Homotopic callosal inhibition. *Brain and Language, 23*(1), 116–125.

Corballis, M. C. (1996). Hemispheric interactions in temporal judgments about spatially separated stimuli. *Neuropsychology, 10*(1), 42–50.

Corballis, M. C. (1998). Sperry and the age of Aquarius: Science, values and the split brain. *Neuropsychologia, 36*(10), 1083–1087.

Corballis, M. C., Boyd, L., Schulze, A., et al. (1998). Role of the commissures in interhemispheric temporal judgments. *Neuropsychology, 12*(4), 519–525.

Crow, T. J., Crow, L. R., Done, D. J., et al. (1998). Relative hand skill predicts academic ability: Global deficits at the point of hemispheric indecision. *Neuropsychologia, 36*(12), 1275–1282.

Csermely, D. (2004). Lateralisation in birds of prey: Adaptive and phylogenetic considerations. *Behavioural Processes, 67*(3), 511–520.

Cummings, J. L. (1997). Neuropsychiatric manifestations of right hemisphere lesions. *Brain and Language, 57*(1), 22–37.

Cutting, J. (1997). *Principles of psychopathology*. Oxford: Oxford University Press.

Damasio, H., Grabowski, T. J., Tranel, D., et al. (1996). A neural basis for lexical retrieval. *Nature, 380*(6574), 499–505 (see comments) (erratum appears in Nature, 381(6595), p. 810).

Denenberg, V. H., Garbanati, J., Sherman, D. A., et al. (1978). Infantile stimulation induces brain lateralization in rats. *Science, 201*(4361), 1150–1152.

Dharmaretnam, M., & Andrew, R. J. (1994). Age-specific and stimulus-specific use of right and left eyes by the domestic chick. *Animal Behaviour, 48*(6), 1395–1406.

Dharmaretnam, M., & Rogers, L. J. (2005). Hemispheric specialization and dual processing in strongly versus weakly lateralized chicks. *Behavioural Brain Research, 162*(1), 62–70.

Dijksterhuis, A., Aarts, H., Bargh, J. A., et al. (2000). Past contact, stereotype strength, and automatic behavior. *Journal of Experimental Social Psychology, 36*, 531–544.

Drake, R. A., & Bingham, B. R. (1985). Induced lateral orientation and persuasibility. *Brain and Cognition, 4*(2), 156–164.

Evans, C. S., Evans, L., & Marler, P. (1993). On the meaning of alarm calls – functional reference in an avian vocal system. *Animal Behaviour, 46*(1), 23–38.

Evert, D. L., & Kmen, M. (2003). Hemispheric asymmetries for global and local processing as a function of stimulus exposure duration. *Brain and Cognition, 51*(1), 115–142.

Eviatar, Z., & Just, M. A. (2006). Brain correlates of discourse processing: An fMRI investigation of irony and conventional metaphor comprehension. *Neuropsychologia, 44*(12), 2348–2359.

Fabre-Thorpe, M., Fagot, J., Lorincz, E., et al. (1993). Laterality in cats: Paw preference and performance in a visuomotor activity. *Cortex, 29*(1), 15–24.

Faust, M., & Mashal, N. (2007). The role of the right cerebral hemisphere in processing novel metaphoric expressions taken from poetry: A divided visual field study. *Neuropsychologia, 45*(4), 860–870.

Federmeier, K. D., & Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research, 8*(3), 373–392.

Feinberg, T. E., Roane, D. M., Kwan, P. C., et al. (1994). Anosognosia and visuoverbal confabulation. *Archives of Neurology, 51*(5), 468–473.

Fernández-Carriba, S., Loeches, A., Morcillo, A., et al. (2002). Asymmetry in facial expression of emotions by chimpanzees. *Neuropsychologia, 40*(9), 1523–1533.

Fink, G. R., Marshall, J. C., Halligan, P. W., et al. (1999). Hemispheric asymmetries in global/local processing are modulated by perceptual salience. *Neuropsychologia, 37*(1), 31–40.

Foldi, N. S. (1987). Appreciation of pragmatic interpretations of indirect commands: Comparison of right and left hemisphere brain-damaged patients. *Brain and Language, 31*(1), 88–108.

Foldi, N. S., Cicone, M., & Gardner, H. (1983). Pragmatic aspects of communication in brain-damaged patients. In S. J. Segalowitz (Ed.), *Language functions and brain organisation* (pp. 51–86). New York: Academic.

Gainotti, G. (2002). The relationships between anatomical and cognitive locus of lesion in category-specific disorders. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 403–426). Hove, UK: Psychology Press.

Gardner, H. (1974). *The shattered mind*. New York: Knopf.

Gazzaniga, M. S., & LeDoux, J. E. (1978). *The integrated mind*. New York: Plenum Press.

Gloning, I., Gloning, K., & Hoff, H. (1968). *Neuropsychological symptoms and syndromes in lesions of the occipital lobe and the adjacent areas*. Paris: Gauthier-Villars.

Goldberg, E. (1990). Associative agnosias and the functions of the left hemisphere. *Journal of Clinical and Experimental Neuropsychology, 12*(4), 467–484.

Goldberg, E., & Costa, L. D. (1981). Hemispheric differences in the acquisition and use of descriptive systems. *Brain and Language, 14*(1), 144–173.

Grossman, M. (1988). Drawing deficits in brain-damaged patients' freehand pictures. *Brain and Cognition, 8*(2), 189–205.

Güntürkün, O., Diekamp, B., Manns, M., et al. (2000). Asymmetry pays: Visual lateralization improves discrimination success in pigeons. *Current Biology, 10*(17), 1079–1081.

Haggard, M. P., & Parkinson, A. M. (1971). Stimulus and task factors as determinants of ear advantages. *Quarterly Journal of Experimental Psychology, 23*(2), 168–177.

Halpern, M. E., Güntürkün, O., Hopkins, W. D., et al. (2005). Lateralization of the vertebrate brain: Taking the side of model systems. *Journal of Neuroscience, 25*(45), 10351–10357.

Hécaen, H., & Albert, M. L. (1978). *Human neuropsychology*. New York: Wiley.

Heilman, K. M., Scholes, R., & Watson, R. T. (1975). Auditory affective agnosia: Disturbed comprehension of affective speech. *Journal of Neurology, Neurosurgery, and Psychiatry, 38* (1), 69–72.

Henson, R. N. A., Rugg, M. D., Shallice, T., et al. (2000). Confidence in recognition memory for words: Dissociating right prefrontal roles in episodic retrieval. *Journal of Cognitive Neuroscience, 12*(6), 913–923.

Hoffman, A. M., Robakiewicz, P. E., Tuttle, E. M., et al. (2006). Behavioural lateralisation in the Australian magpie (Gymnorhina tibicen). *Laterality, 11*(2), 110–121.

Hopkins, W. D., & Marino, L. (2000). Asymmetries in cerebral width in nonhuman primate brains as revealed by magnetic resonance imaging (MRI). *Neuropsychologia, 38*(4), 493–499.

Höppner, J., Kunesch, E., Buchmann, J., et al. (1999). Demyelination and axonal degeneration in corpus callosum assessed by analysis of transcallosally mediated inhibition in multiple sclerosis. *Clinical Neurophysiology, 110*(4), 748–756.

Hoptman, M. J., & Davidson, R. J. (1994). How and why do the two cerebral hemispheres interact? *Psychological Bulletin, 116*(2), 195–219.

Hough, M. S. (1990). Narrative comprehension in adults with right and left hemisphere brain-damage: Theme organization. *Brain and Language, 38*(2), 253–277.

Ivry, R. B., & Robertson, L. C. (1998). *The two sides of perception*. Cambridge, MA: Massachusetts Institute of Technology.

Jäncke, L., & Steinmetz, H. (2003). Anatomical brain asymmetries and their relevance for functional asymmetries. In K. Hugdahl & R. J. Davidson (Eds.), *The asymmetrical brain* (pp. 187–230). Cambridge, MA: Massachusetts Institute of Technology Press.

Kaplan, J. A., Brownell, H. H., Jacobs, J. R., et al. (1990). The effects of right hemisphere damage on the pragmatic interpretation of conversational remarks. *Brain and Language, 38*(2), 315–333.

Kimura, D. (1963). Right temporal-lobe damage: Perception of unfamiliar stimuli after damage. *Archives of Neurology, 8*, 264–271.

Kinsbourne, M. (1988). Hemispheric interactions in depression. In Kinsbourne, M. (Ed.), *Cerebral hemisphere function in depression* (pp. 133–162). Washington, DC: American Psychiatric Press.

Kinsbourne, M. (1993). Orientational bias model of unilateral neglect: Evidence from attentional gradients within hemispace. In I. H. Robertson & J. C. Marshall (Eds.), *Unilateral neglect: Clinical and experimental studies* (pp. 63–86). Hove: Lawrence Erlbaum.

Kitterle, F. L., Christman, S., & Hellige, J. B. (1990). Hemispheric differences are found in the identification, but not the detection, of low versus high spatial frequencies. *Perception and Psychophysics, 48*(4), 297–306.

Kosslyn, S. M. (1987). Seeing and imagining in the cerebral hemispheres: A computational approach. *Psychological Review, 94*(2), 148–175.

Landis, T., Cummings, J. L., Benson, D. F., et al. (1986). Loss of topographic familiarity: An environmental agnosia. *Archives of Neurology, 43*(2), 132–136.

Lippolis, G., Bisazza, A., Rogers, L. J., et al. (2002). Lateralisation of predator avoidance responses in three species of toads. *Laterality, 7*(2), 163–183.

Lippolis, G., Westman, W., McAllan, B. M., et al. (2005). Lateralisation of escape responses in the striped-faced dunnart, Sminthopsis macroura (*Dasyuridae*: *Marsupialia*). *Laterality, 10*(5), 457–470.

Locke, J. (1690). *An essay on human understanding*. London: Rivington.

Marsolek, C. J. (1995). Abstract visual-form representations in the left cerebral hemisphere. *Journal of Experimental Psychology: Human Perceptual Performance, 21*(2), 375–386.

Martin, A., Wiggs, C. L., Ungerleider, L. G., et al. (1996). Neural correlates of category-specific knowledge. *Nature, 379*(6566), 649–652.

Marzi, C. A., Bisiacchi, P., & Nicoletti, R. (1991). Is interhemispheric transfer of visuomotor information asymmetric? Evidence from a meta-analysis. *Neuropsychologia, 29*(12), 1163–1177.

Mashal, N., & Faust, M. (2008). Right hemisphere sensitivity to novel metaphoric relations: Application of the signal detection theory. *Brain and Language, 104*(2), 103–112.

Mashal, N., Faust, M., & Hendler, T. (2005). The role of the right hemisphere in processing nonsalient metaphorical meanings: Application of principal components analysis to fMRI data. *Neuropsychologia, 43*(14), 2084–2100.

Mashal, N., Faust, M., Hendler, T., et al. (2007). An fMRI investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain and Language, 100*(2), 115–126.

McGilchrist, I. (2009). *The master and his emissary: The divided brain and the making of the Western World*. New Haven: Yale University Press.

McGrew, W. C., & Marchant, L. F. (1999). Laterality of hand use pays off in foraging success for wild chimpanzees. *Primates, 40*(3), 509–513.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Meador, K. J., Loring, D. W., Feinberg, T. E., et al. (2000). Anosognosia and asomatognosia during intracarotid amobarbital inactivation. *Neurology, 55*(6), 816–820.

Meyer, B.-U., Röricht, S., Gräfin von Einsiedel, H., et al. (1995). Inhibitory and excitatory interhemispheric transfers between motor cortical areas in normal subjects and patients with abnormalities of the corpus callosum. *Brain, 118*(2), 429–440.

Michel, E. M., & Troost, B. T. (1980). Palinopsia: Cerebral localization with CT. *Neurology, 30*(8), 887–889.

Mills, L., & Rollman, G. B. (1980). Hemispheric asymmetry for auditory perception of temporal order. *Neuropsychologia, 18*(1), 41–48.

Milo, R. G., & Quiatt, D. (1993). Glottogenesis and anatomically modern Homo sapiens: The evidence for and implications of a late origin of vocal language. *Current Anthropology, 34*(5), 569–598.

Mithen, S. J. (1998). A creative explosion? Theory of mind, language and the disembodied mind of the upper Palaeolithic. In S. J. Mithen (Ed.), *Creativity in human evolution and prehistory* (pp. 165–192). London: Routledge & Kegan Paul.

Mithen, S. J. (2005). *Singing Neanderthals: The origin of music, language, mind and body*. London: Phoenix.

Müller, T., Büttner, T., Kuhn, W., et al. (1995). Palinopsia as sensory epileptic phenomenon. *Acta Neurologica Scandinavica, 91*(6), 433–436.

Mummery, C. J., Patterson, K., Hodges, J. R., et al. (1996). Generating “tiger” as an animal name or a word beginning with T: Differences in brain activation. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 263*(1373), 989–995.

Nagel, T. (1986). *The view from nowhere*. Oxford: Oxford University Press.

Nebes, R. D. (1974). Hemispheric specialization in commissurotomized man. *Psychological Bulletin, 81*(1), 1–14.

Nicholls, M. E. R. (1994). Hemispheric asymmetries for temporal resolution: A signal detection analysis of threshold and bias. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 47*(2), 291–310.

Oliveri, M., Rossini, P. M., Traversa, R., et al. (1999). Left frontal transcranial magnetic stimulation reduces contralesional extinction in patients with unilateral right brain damage. *Brain, 122*(9), 1731–1739.

Ornstein, R., Herron, J., Johnstone, J., et al. (1979). Differential right hemisphere involvement in two reading tasks. *Psychophysiology, 16*(4), 398–401.

Perani, D., Cappa, S. F., Bettinardi, V., et al. (1995). Different neural systems for the recognition of animals and man-made tools. *NeuroReport, 6*(12), 1637–1641.

Phelps, E. A., & Gazzaniga, M. S. (1992). Hemispheric differences in mnemonic processing: The effects of left hemisphere interpretation. *Neuropsychologia, 30*(3), 293–297.

Price, C. J., & Friston, K. J. (2002). Functional imaging studies of category specificity. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 427–447). Hove: Psychology Press.

Rapp, A. M., Leube, D. T., Erb, M., et al. (2004). Neural correlates of metaphor processing. *Brain Research. Cognitive Brain Research, 20*(3), 395–402.

Rausch, R. (1977). Cognitive strategies in patients with unilateral temporal lobe excisions. *Neuropsychologia, 15*(3), 385–335.

Rausch, R. (1985). Differences in cognitive function with left and right temporal lobe dysfunction. In D. F. Benson & E. Zaidel (Eds.), *The dual brain* (pp. 247–261). New York: Guilford.

Rizzolatti, G., Fogassi, L., Gallese, V., et al. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*(9), 661–670.

Robertson, L. C., Lamb, M. R., & Knight, R. T. (1988). Effects of lesions of temporal-parietal junction on perceptual and attentional processing in humans. *Journal of Neuroscience, 8*(10), 3757–3769.

Robertson, L. C., & Lamb, M. R. (1991). Neuropsychological contributions to theories of part/whole organisation. *Cognitive Psychology, 23*, 299–330.

Robins, A., & Rogers, L. J. (2006). Complementary and lateralized forms of processing in Bufo marinus for novel and familiar prey. *Neurobiology of Learning and Memory, 86*(2), 214–227.

Rogers, L. J. (2000). Evolution of hemisphere specialisation: Advantages and disadvantages. *Brain and Language, 73*(2), 236–253.

Rogers, L. J. (2005). Cognitive and social advantages of having a lateralized brain. In Y. B. Malashichev & A. W. Deckel (Eds.), *Behavioral and morphological asymmetries in vertebrates* (pp. 129–139). Austin: Landes Bioscience.

Rogers, L. J., & Kaplan, G. (2006). An eye for a predator: Lateralization in birds, with particular reference to the Australian magpie. In Y. B. Malashichev & A. W. Deckel (Eds.), *Behavioral and morphological asymmetries in vertebrates* (pp. 47–57). Austin: Landes Bioscience.

Rogers, L. J., & Workman, L. (1989). Light exposure during incubation affects competitive behaviour in domestic chicks. *Applied Animal Behaviour Science, 23*, 187–198.

Rogers, L. J., Zucca, P., & Vallortigara, G. (2004). Advantages of having a lateralized brain. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 271*(Suppl. 6), S420–S422.

Röricht, S., Irlbacher, K., Petrow, E., et al. (1997). Normwerte transkallosal und kortikospinal vermittelter Effekte einer hemisphärenselektiven magnetischen Kortexreizung beim

Menschen. *Zeitschrift für Elektroenzephalographie, Elektromyographie und Verwandte Gebiete, 28*, 34–38.

Ross, E. D., Thompson, R. D., & Yenkosky, J. (1977). Lateralisation of affective prosody in the brain and callosal integration of hemispheric language functions. *Brain and Language, 56*(1), 27–54.

Saron, C. D., & Davidson, R. J. (1989). Visual evoked potential measures of interhemispheric transfer time in humans. *Behavioral Neuroscience, 103*(5), 1115–1138.

Saron, C. D., Foxe, J. J., Simpson, G. V., et al. (2002). Interhemispheric visuomotor activation: Spatiotemporal electrophysiology related to reaction time. In E. Zaidel & M. Iacoboni (Eds.), *The parallel brain: The cognitive neuroscience of the corpus callosum* (pp. 171–219). Cambridge, MA: Massachusetts Institute of Technology Press.

Saron, C. D., Foxe, J. J., Schroeder, C. E., et al. (2003). Complexities of interhemispheric communication in sensorimotor tasks revealed by high-density event-related potential mapping. In K. Hugdahl & R. J. Davidson (Eds.), *The asymmetrical brain* (pp. 341–408). Cambridge, MA: Massachusetts Institute of Technology Press.

Schmidt, G. L., DeBuse, C. J., & Seger, C. A. (2007). Right hemisphere metaphor processing? Characterizing the lateralization of semantic processes. *Brain and Language, 100*(2), 127–141.

Schneiderman, E. I., Murasugi, K. G., & Saddy, J. D. (1992). Story arrangement ability in right-brain damaged patients. *Brain and Language, 43*(1), 107–120.

Schnider, A., Gutbrod, K., Hess, C. W., et al. (1996). Memory without context: Amnesia with confabulations after infarction of the right capsular genu. *Journal of Neurology, Neurosurgery, and Psychiatry, 61*(2), 186–193.

Sergent, J. (1982). The cerebral balance of power: Confrontation or cooperation? *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 253–272.

Starkstein, S. E., Fedoroff, J. P., Price, T. R., et al. (1992). Anosognosia in patients with cerebrovascular lesions: A study of causative factors. *Stroke, 23*(10), 1446–1453.

Stringaris, A. K., Medford, N. C., Giampietro, V., et al. (2007). Deriving meaning: Distinct neural mechanisms for metaphoric, literal, and non-meaningful sentences. *Brain and Language, 100*(2), 150–162.

Stump, E. (2009). The problem of evil: Analytic philosophy and narrative. In O. D. Crisp & M. C. Rea (Eds.), *Analytic theology: New essays in the philosophy of theology* (pp. 251–264). New York: Oxford University Press.

Stuss, D. T. (1991). Disturbance of self-awareness after frontal system damage. In G. P. Prigatano & D. L. Schacter (Eds.), *Awareness of deficit after brain injury: Clinical and theoretical issues* (pp. 66–83). Oxford: Oxford University Press.

Swisher, L., & Hirsh, I. J. (1972). Brain damage and the ordering of two temporally successive stimuli. *Neuropsychologia, 10*(2), 137–152.

Tootell, R. B., Mendola, J. D., Hadjikhani, N. K., et al. (1998). The representation of the ipsilateral visual field in human cerebral cortex. *Proceedings of the National Academy of Sciences of the USA, 95*(3), 818–824.

Vallortigara, G. (1992). Right hemisphere advantage for social recognition in the chick. *Neuropsychologia, 30*(9), 761–768.

Vallortigara, G., Rogers, L. J., Bisazza, A., et al. (1998). Complementary right and left hemifield use for predatory and agonistic behaviour in toads. *NeuroReport, 9*(14), 3341–3344.

van Kleeck, M. H. (1989). Hemispheric differences in global versus local processing of hierarchical visual stimuli by normal subjects: New data and a meta-analysis of previous studies. *Neuropsychologia, 27*(9), 1165–1178.

Ventolini, N., Ferrero, E. A., Sponza, S., et al. (2005). Laterality in the wild: Preferential hemifield use during predatory and sexual behaviour in the black-winged stilt (*Himantopus himantopus*). *Animal Behaviour, 69*, 1077–1084.

Vogeley, K., Bussfeld, P., Newen, A., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1, Pt 1), 170–181.

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain, 110*(5), 1273–1296.
Wymer, J. H., Lindman, L. S., & Booksh, R. L. (2002). A neuropsychological perspective of aprosody: Features, function, assessment, and treatment. *Applied Neuropsychology, 9*(1), 37–47.
Yoshida, T., Yoshino, A., Takahashi, Y., et al. (2007). Comparison of hemispheric asymmetry in global and local information processing and interference in divided and selective attention using spatial frequency filters. *Experimental Brain Research, 181*(3), 519–529.

Neurotheological Eudaimonia 101

Andrew Pinsent

Contents

Abstract
The Aristotelian concept of *eudaimonia*, variously translated as "happiness" or "flourishing," has been a foundational principle of the Western philosophical tradition of virtue ethics. The theology of Thomas Aquinas transfigured the meaning of *eudaimonia*, introducing the concept of "infused" virtues. Rather than being acquired through habituation, these virtues can be understood metaphorically as removing a person's "spiritual autism," enabling second-person relatedness to God. This chapter applies these general principles to examine the everyday role of second-person relatedness in ethical formation, reviewing research into the neural conditions and concomitants of joint attention, and implications for a new understanding of virtue ethics.

Introduction

Anyone who doubts that what is denoted by religious language is widely perceived as having at least some connection to happiness should take a cursory glance at the products of the modern advertising industry. Promotional messages are replete with

A. Pinsent
Ian Ramsey Centre for Science and Religion, University of Oxford, Oxford, UK
e-mail: andrew.pinsent@hmc.ox.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_100,

words such as "heaven," "paradise," "bliss," and "nirvana," the latter presumably being interpreted by Western minds as heaven with a hint of Eastern promise. Advertisers appear to judge that transient goods, such as health foods or holidays, become more desirable when associated with vaguely religious words, especially those that denote a state of eternal happiness that has proved elusive to unaided human efforts but that features as a goal of many kinds of religious practice. Hence, it is not inappropriate that, in this handbook, the chapter on neurotheology, which applies neuroscience to examine aspects of ultimate human goals within a theological perspective, should be placed close to the section on neuromarketing.

Any study of happiness is complicated by well-known ambiguities, however, such as the need to distinguish what a person seeks subjectively, depending at least in part on her personal history and the "desires of her heart" (Stump 2010) and what is good for the person objectively, a distinction to which advertisers are not always eager to draw attention. Rather than the word "happiness," this chapter therefore adopts the term *eudaimonia*, sometimes translated as "flourishing," taken directly from the *Nicomachean Ethics* (NE) of Aristotle, the "canonical text" (MacIntyre 2007) of the Western philosophical tradition regarding such matters. Besides denoting what everyone seeks, albeit inchoately, the term *eudaimonia* also conveys a sense of objective excellence, especially of what is specific to human life as compared to that of animals and plants (NE I.7, 1097b33-1098a7). For Aristotle, what is most unique about human beings in comparison to plants and animals is their rational principle. Hence his account of *eudaimonia* in the *Nicomachean Ethics* centers on intellectual activity in accordance with virtue or excellence in a complete life (NE I.4, 1098a15-19). This *eudaimonia* is not associated with any specific religious revelation. Nevertheless, it would be misleading to claim that there is no role for the divine, since God is treated elsewhere in Aristotle's work as a first cause and hence as an object, perhaps even the supreme object, of philosophical enquiry in a flourishing life (*Metaphysics* I.1, 983a5-10; XII, 7, 1072b25-30). Moreover, the etymology of the word "*eudaimon*" implies "living in a way that is well-favored by a god" (Kraut 2012), and Aristotle thought that the gods, if they give any gifts to human beings, might be expected to bestow whatever is required for *eudaimonia* on those who strive to be most like themselves, such as philosophers (NE X.8, 1179a22-32).

"Theological *eudaimonia*," as the term is used in this chapter, follows Aristotle's general approach to *eudaimonia* as an activity flowing from virtue, but within the framework of a revealed relationship to the divine that culminates in a state that Aristotle claimed to be impossible, namely, friendship with God (NE VIII.7, 1159a3-5). Specifically, this chapter examines findings in contemporary neuroscience that appear to offer insights pertaining to a new metaphoric understanding of the work of Thomas Aquinas, who developed in the thirteenth century an extraordinarily detailed and influential transformation of Aristotelian virtue ethics. Strictly speaking, this examination is of those embodied experiences that seem to provide the closest everyday parallels to Aquinas's descriptions of theological *eudaimonia*, rather than the purported relationship to God directly. Nevertheless, studies of these parallels are important to evaluate their validity and explore their implications. Moreover, these implications go beyond the theological framework within which

Aquinas's approach was developed, suggesting a new, non-Aristotelian approach to human flourishing generally, one to which contemporary neuroscience is well placed to contribute. Prior to examining the neuroscience, however, it is necessary to review briefly Aquinas's account of *eudaimonia* within his own theological framework and how this differs from the classical and more familiar account of the Western philosophical tradition.

Aquinas on Theological *Eudaimonia*

Aquinas's presents his most detailed account of theological *eudaimonia* in his largest and most famous work, the *Summa theologiae* (ST), which devotes 1,004 articles to the virtues and associated matters, including a generic overview of the various categories in ST 1a2ae qq.55-89 and a systematic account of specific virtues in ST 2a2ae qq.1-170. Much pertinent and insightful material can be found elsewhere, and his scriptural commentaries can be helpful in providing an interpersonal and holistic counterbalance to the fine-grained thematic analysis of many other texts. Nevertheless, Aquinas's mature understanding of the characteristic dispositions of a life that looks toward friendship with God is to be found in all its essential details in the ST.

For many reasons, including Aquinas's admiration for Aristotle, the parallels in their respective accounts, and the pervasive influence of the *Nicomachean Ethics* upon the history of virtue ethics generally, it has long been assumed that Aristotle is the "dominant" though not exclusive voice in Aquinas's treatment of the virtues (McInerny 1993). In recent years, however, a growing body of theologians and analytic philosophers have raised questions about the validity of this interpretation. An indication of the extent of divergence from Aristotle can be seen by comparing the structure of Aquinas's most detailed systematic account of particular virtues (ST 2a2ae qq.1-170) with its counterpart in the *Nicomachean Ethics*. Preceding the four cardinal virtues, the names of which were familiar in pagan antiquity, Aquinas introduces the so-called theological virtues of faith, hope, and "love" or divine friendship (qq.1-46) to make seven major virtues in total. Yet even those virtues that at first seem familiar from classical accounts, such as prudence, justice, temperance, and courage, acquire novelties in the ST. For example, Aquinas adds virtues that are rare in classical writings, such as humility (q.161), and incorporates a great diversity of questions and priorities that are unknown to the *Nicomachean Ethics*. Such issues include, for example, whether backbiting is gravely evil (q.73), whether joy is an effect of devotion (q.82), and whether God should be praised in song (q.91).

At a more fundamental level, however, even the very notion of a virtue in the ST diverges radically from Aristotle's account in ways to which Eleonore Stump, for instance, has drawn attention (Stump 2011). To give a particularly important example, a perfect or true virtue is infused by God rather than acquired by practice according to Aquinas (ST 1a2ae q.55, a.4; q.65, a.3). These infused virtues are unified by the theological virtue of love and can be lost, in the sense of being "cut-off" as virtues, or restored, by singular actions of grave evil or reconciliation rather than by gradual habituation (ST 1a2ae q.71 a.4). What is also puzzling is that

Aquinas's account of these infused virtues is interwoven with other perfective dispositions that are not virtues at all, but dispositions called "gifts" (ST 1a2ae, q.68). According to Aquinas, the gifts are also infused and pertain to the same things as the virtues, but operate in a different manner (ST 1a2ae, q.68, a.4). Whereas a virtue disposes a person to be moved by her own reason, a gift is triadic, disposing a person to be moved by God with respect to some object, in what Aquinas describes as a "union of the soul" with God (ST 1a2ae, q.68, a.1; 2a2ae, q.45, a.3).

Aquinas's account of these virtues and gifts has long proved mysterious, and it is unsatisfactory simply to draw up a catalog of differences from the *Nicomachean Ethics*. The challenge, as Jean Porter has explained, is that "the infused virtues function in a way that is significantly different from the way in which the acquired virtues function, so much so that they can be described as virtues only in a carefully qualified sense" (Porter 1992). The deeper problem is that if language is to have any agreed meaning, this meaning plausibly depends on being able to associate words with embodied experience by means of metaphor, an association given some additional credence from recent work in neuroscience (McGilchrist 2009). In the Western philosophical tradition, however, Aristotle has associated the term "virtue" with the experience of habituation, comparable to practicing a sport or musical instrument (NE II.1, 1103b6-22), an experience that is irreconcilable with Aquinas's claims about the infused virtues. If it is possible to understand properly what Aquinas means by an infused virtue, an alternative embodied experience is required. Based on Aquinas's detailed descriptions of the virtues and gifts, it is arguable that the appropriate experience is what experimental psychologists today call "joint attention" (Pinsent 2012), an aspect of social cognition that involves a "sharing an awareness of the sharing of the focus" on some object with another person, something that often entails "sharing an attitude towards the thing or event in question" (Hobson 2005). The infused dispositions that Aquinas describes can be understood as putting a person into a joint attention relationship with God. This joint attention relationship forms a person's attitudes to all kinds of other matters in ways that will often be different to what these attitudes would be in the absence of this relationship.

From this interpretation, there is another metaphor for the effect of the infused dispositions described by Aquinas. Some scholars have suggested that joint attention may be the root experience for understanding what it means to relate to a person specifically as "I" to "you," and for expressing such relatedness in language (Heal 2005; Roessler 2005). Moreover, the study of autistic spectrum disorder has given some empirical weight to this association, insofar as a diagnosis of autism in children is often associated both with a deficiency of engagement in joint attention (Wimpory et al. 2000) and with difficulties in learning the correct use of second-person forms of grammar (Tager-Flusberg 1994). On this basis, another way of understanding Aquinas's account of the infused dispositions is that they enable specifically second-person relatedness to God, removing what could be likened to an innate *spiritual autism* of unaided human nature from the divine perspective. From this perspective, the question the Lord God directs to Adam after the narrative of the Fall in the Garden of Eden, "Where are you?" (Gn 3:9), and perhaps also in part the set of rhetorical questions directed by the Lord to Job,

"Where were you?" (Jb 38-41), articulates a loss of second-person relatedness with human beings that the infused virtues and gifts of theological *eudaimonia* are intended to restore.

This interpretation in terms of second-person relatedness should of course be treated as a metaphor, like the removal of physical blindness has long been used as a metaphor for intellectual enlightenment. Nevertheless, this relational metaphor has considerable explanatory power in unifying what would otherwise appear as rather peculiar and ad hoc claims about theological *eudaimonia*. For example, as noted previously, Aquinas claims that the infused virtues, unified by the theological virtue of love, can be lost, in the sense of being "cut-off" as virtues, or restored, by singular actions of grave evil or reconciliation, a claim that makes no sense at all from a traditional understanding of virtue in terms of habituation. If, however, the goal of the virtues is divine friendship, a state of harmonized second-person relatedness, then it is easier to see how a relationship can be betrayed or restored by a single action. Following a betrayal and concomitant loss of relationship, acquired good habits of daily life will not suddenly vanish, but they will cease to be virtues that are able to contribute to a state of second-personal flourishing with the other person, at least until the relationship is restored (Pinsent 2012). For this reason, wrongdoing in this worldview is also "sin," with the connotations of betraying a second-person relationship that is often described using the language of "covenant" rather than contract.

The account presented above is a theological one, and presupposes the existence of supernatural elements such as a personal and loving God who is willing and able to infuse dispositions into persons that enable them to relate to God in a second-personal way, culminating in friendship. Nevertheless, it is not difficult to see that the general framework of Aquinas's account has far broader implications. In everyday life, the development of virtues may be dependent on second-person relatedness in ways that are opaque to a traditional Aristotelian analysis. For example, a commonplace experience is that children first acquire temperate behavior with respect to food and drink not through selecting a virtuous mean established by practical wisdom, but generally by some kind of second-person interaction, such as a game, with a parent or caregiver (Werle et al. 1993). Such considerations highlight the need to examine further the association of second-person relatedness and character formation and, in particular, to examine what neuroscience can contribute.

Second-Person Relatedness and Social Neuroscience

Aquinas's account of how a person's behavior is shaped by a relationship to a personal God is radical as much for what it excludes as for what it includes. There is no trace of reasoning along the lines of the following, "I desire *P* in order to flourish. If I do *Q*, God will grant me *P*; therefore I do *Q*." More generally, God is not simply a means to an end and does not enter into Aquinas's account simply as some all-powerful third-personal or impersonal agent capable of adding or subtracting various goods of human flourishing. Instead, true flourishing consists

in a perfection of second-person relatedness, with God culminating in divine friendship. Across the spectrum of what is encompassed by the field of social neuroscience, what is therefore of greatest relevance to this account of flourishing is whatever pertains specifically to second-person relatedness in ordinary and everyday human relations. As noted previously, such relatedness is not simply awareness of a person or interpersonal interaction generally, but has been described as a shared awareness of the sharing of focus with another person, which often involves sharing an attitude toward the object of shared attention. The object of the attention may be a physical object, a shared activity or possibly, by perceived cues, an abstract idea. Such an interaction may be almost too slight to register, as in a momentary glance of recognition or a smile between strangers, or profound and sustained with a lifelong friend. Nevertheless, across the range of possible situations, the experience of joint attention implies a sense of the other person being, in some sense, present to oneself, a presence that does not appear necessary when simply thinking about someone, or thinking about what another person is thinking about. Moreover, the subjective experience of joint attention is typically described as a "meeting of minds" with the other person (Eilan 2005).

Framed in these terms, it is possible to begin to sift the most relevant findings available from social neuroscience, but caution is needed in identifying and describing the most appropriate phenomena. For a start, there is an influential school of thought in contemporary philosophy that holds that the subjective experience of a thing is irreducible to knowledge about a thing, regardless of the detail of that knowledge (Nagel 1974), and experience of a person as a person is plausibly among those things that are most irreducible. In social neuroscience, however, this distinction sometimes risks being obscured a priori by much of the language in which knowledge of persons is expressed. For example, it is commonplace to describe the way in which one being-with-a-mind identifies and relates to another being-with-a-mind in terms of "Theory of Mind" (ToM), reference to which was made in 1 % of all academic publications in psychology that referred to infants or children in 2003 (Reddy and Morris 2004). Although it is true that ToM may be interpreted as "more like a loose coalition than a unified movement," the mere label of a field of research or a "folk psychological" concept (Costall and Leudar 2004), the word "theory" almost invariably invokes in us embodied experiences connected with simplified, formal, and quantitative representations of the world that are inherently objective rather than subjective. Moreover, the term "theory" is typically employed to describe situations in which what is being theorized about is objective and does not need to be personally present. Even if the term in fact turns out to be an appropriate one, it therefore seems prudent to focus on those interpersonal phenomena that seem at first to be most *unlike* one person theorizing about the mind of another, if the aim is to discover what, if anything, is most distinctive about second-person relatedness.

The remainder of this section therefore focuses on those aspects of social neuroscience involving interpersonal relatedness in situations of mutual presence that, for one reason or another, appear closer to "thinking with" rather than "theorizing about" the other person. To give some examples to clarify this

emphasis, Colwyn Trevarthen has noted the many ways in which a responsive mother "dovetails" with her infant (Trevarthen 1979), and musicians playing in harmony in an orchestra also seem to be thinking with rather than theorizing about one another. Something like this distinction has also made by Hein and Singer's proposed two pathways for understanding another person's mind, namely, empathy and embodiment of another's emotions versus theory of mind, mentalizing, or cognitive perspective taking. Moreover, they also claim that these distinct modes of understanding involve activation of different areas of the brain (Hein and Singer 2008), although the extent to which a precise differentiation is possible or meaningful is a matter of debate (Adolphs and Janowski 2011). What can be said, however, is there is growing evidence of distinct modes of understanding of the minds of others, at least one of which can be described broadly as thinking with a person.

So what kinds of phenomena in social neuroscience are most appropriate to this "thinking with" another person in second-person relatedness, especially insofar as the latter experience, may serve as a metaphor for theological *eudaimonia*? Philosophers from at least the time of Aristotle have highlighted how any intellectual operations that give rise to human action must have cognitive and affective aspects. Such aspects may be practically synchronous and inseparable, with the affective aspects being far easier to detect than cognition alone by means of behavioral or physiological changes. The famous "Eureka!" moment attributed to Archimedes would not have been so dramatically obvious to his fellow citizens if he had merely understood the answer to the problem set him by Hiero of Syracuse without also expressing his joy in an unconventional way. Moreover, proponents of what is sometimes called "common coding theory" have emphasized how some of the neurological correlates of thought and action may be practically inseparable (Sperry 1952). Certainly, the direction of causation between thought and action is not wholly one way, shown by the way in which the disruption of human premotor cortex impairs speech perception (Meister et al. 2007) and a wide range of research showing how mental simulations take place when processing information (Niedenthal et al. 2011). Furthermore, consideration of the many human actions that happen without conscious reflection, for example, swift responses to sudden threats, or habitually trained "lower level" actions such as the movement of fingers on a keyboard, along with much neuropsychological research on automatic brain responses, underlines how the mental processes that give rise to human action are much broader than those we reflect on consciously (Libet 1985). Nevertheless, if intellectual operation, whether conscious or not, is properly a cause of any human action, it is hard to circumvent the need for such operations to have cognitive and affective aspects, respectively, the cognition of a thing and attitude toward the thing as good to pursue or avoid. Moreover, recent research has suggested that cognitive and affective aspects of mental activity are even dissociable under certain conditions (Shamay-Tsoory and Aharon-Peretz 2007; Völlm et al. 2006; Hynes et al. 2006). Hence also one would expect "thinking with" operations of second-person relatedness, if they are to be the basis of any human action, to have both cognitive and affective aspects.

Second-person relatedness, however, is complicated by the fact that "thinking with" situations have a triadic person-person-object form. Hence, if there are dedicated processes in mental activity that are attuned to this kind of relatedness, one would expect some such processes to be directed to the object of shared attention, with others dedicated to establishing, monitoring, and maintaining union with the second person. Such processes would divide roughly along the following lines: (1) cognition of a second person; (2) cognition of harmonization with a second person, and a pre-disposition to favor such harmonization; (3) joint cognition of an object; (4) joint stance toward the object, where "stance" is "a conative attitude prompted by the mind's understanding" (Stump 2011). Note that in actual joint attention, the second person may only enter into the periphery of personal experience. The following paragraphs list some examples of phenomena pertinent to these processes in social neuroscience.

Cognition of a second person: Both the existence and importance of specific second-person cognition are strongly indicated by the rapid ability of newborn infants to identify persons from the range of beings that they first encounter, and by their interest in faces and the imitation of facial expressions (Meltzoff and Moore 1977). On the latter observation, it should be noted that a special connection between face cognition and second-person relatedness has long been drawn by theologians, artists, and philosophers in the tradition of Martin Buber, Emmanuel Levinas, and their successors. More recently, there has been a great variety of accumulating evidence for specific neuronal activity correlating with face cognition, as compared to the cognition of non-facial objects (see, for example, Thompson 1980; Yin 1969; Assal 2001; Bodamer 1947; Farah et al. 1995; Freiwald et al. 2009; Perrett et al. 1985; Rolls 2007). The role of the human face in typical second-person relatedness has also been highlighted by deficits of face cognition in cases of autistic spectrum disorder (Klin et al. 1999) and, conversely, some of the symptoms of autism in cases of prosopagnosia ("face-blindness") (Dalrymple et al. 2012). Face cognition is not the only channel for second-person relatedness, however, as suggested for example by evidence of neural processes that correlate with hearing the sound of human voices (Belin 2011).

Cognition and enjoyment of harmonization with second person: A wide range of evidence suggests that there are neural processes dedicated to the ability to cognize and enjoy harmonization with a second person. The rapid imitation by newborn infants of the facial expressions of adults, noted above, is not only evidence of face-cognition abilities but is also an example of an innate and early disposition toward such harmonization. Moreover, infants as young as 3 months shift their visual attention in the direction toward which the eyes of an adult face turn (Hood et al. 1998). This ability to engage in joint attention appears at roughly the same time, or perhaps even slightly precedes, an infant's first-person ability to flexibly orient attention to objects in her environment (Kirwan et al. 2011). The cognition and enjoyment of harmonization with a second person will typically continue throughout life, often at a level that is subliminal to more overt communications such as speech. To give some of many examples, the "chameleon effect" and the fact that being imitated increases liking (Chartrand and Bargh 1999) may be another

symptom of the way in which we are predisposed toward a harmonization with others. In addition, there is evidence of an innate distaste for situations in which second-person relatedness is not quite right, is interrupted or mistaken. For example, human beings have a negative emotional response to simulations in which something is almost but not quite like a second person, a problem known to animators as the “uncanny valley” (Mori 1970; MacDorman and Ishiguro 2006). What may also be of relevance is the fact that human beings are predisposed to pay a high price for second-person relatedness, even if such relatedness is not fully actualized. A possible example is the way in which many subjects in prisoners’ dilemma games prefer mutual cooperation even though it is in their self-interest to defect whatever the other player does (Kiyonari et al. 2000). A relative lack of mutual cooperation, and a more utilitarian approach to such games, tends to be associated with states in which cooperative inclinations have been inhibited or suppressed, either by brain lesions (Ciaramelli et al. 2007; Koenigs et al. 2007) or possibly by long periods of training in economics (Frank et al. 1993).

Joint cognition of an object: The issues of most conceptual interest under this heading are: (i) whether or not there is a shift in the nature of cognition when shifting from individual attention to joint attention, as opposed to a mere change in external factors such as learning where to direct one’s attention; and (ii) whether joint cognition is in principle reducible to psychological states of an individual which do not already imply such cognition with a second person; or else whether joint cognition is a “primitive phenomenon” (Campbell 2005). In regard to such questions, a promising discovery has been the mirror neurons, which have been directly observed in primate and some other species such as birds, to fire both when the animal acts and when the animal observes the same action performed by another (Di Pellegrino et al. 1992; Gallese et al. 1996; Ferrari et al. 2003). Such operations are often expressed in terms of “simulating” some cognitive process of another being by the activation of parallel cognitive processes that, if carried into action, would produce similar behavior. Even the term “simulation” may be subtly misleading, however, since this word in daily life typically denotes a replication based on abstract models of reality. By contrast, the immediacy of the mirror neuron activation and the fact that such activity is found in a wide spectrum of nonhuman animals suggests that this replication is a primitive phenomenon that does not depend on explicit or effortful reasoning. Although the operation of mirror neurons in human beings cannot be observed directly, there is strong indirect evidence for them in the human brain (Iacoboni 1999). Moreover, the mirror neuron areas in human beings are not simply involved in registering actions, but activate in specific ways upon the observation of intentional actions, such as grasping an object (Iacoboni et al. 2005). Nevertheless, claims that mirror neurons explain “action understanding” probably go beyond what evidence can support (Hickok 2009), not least because of the temptation to mix the language of psychological powers such as “understanding” with those of the neural conditions and concomitants for the exercise of such powers (Bennett and Hacker 2003). Perhaps the best that can be said is there are promising candidates for at least some of the neural conditions and concomitants

required to permit an exercise of joint cognition that is "primitive," both in the sense of being irreducible and of playing a foundational role in early cognitive development.

A joint stance toward the object: There is a wide range of evidence supporting the principle that a stance toward some object of joint attention is different to what the stance would be in the absence of such relatedness, even when it is the object, and not the other person, that is the direct focus of attention. An everyday example is joint attention with infants, in which the infant displays a preferential interest in the object that the caregiver is also attending. Particular tones and melodic contours of what is sometimes called "motherese" plausibly become associated with a particular stance long before infants become sensitive to the segmented words of language (Donald 2001; Falk 2009). The emotional content communicated in this way arguably helps to foster the remarkable ability of human infants to show recognition of intentions from vocal utterances, a point of divergence from apes that also contributes to the acquisition of language (Bloom 2000; Gärdenfors 2003; Kuhl 2011; Tomasello 1998). Yet even ordinary language between adults also involves much more than propositional speech. The distinction and importance of prosody and other nonverbal communication has been shown in many ways, for example, by the effects of damage to the right hemisphere's perisylvian region correlated with an impaired ability to express emotion using prosody (Heilman et al. 2004). Such nonverbal communications, which cannot easily be abstracted from the context of a specific "I" communicating with a specific "you," have reemphasized the need to think about language not simply in terms of objective symbol use and organization, but as a communicative interaction between persons (Nusbaum 2011). These findings are consistent with a general account of nonverbal communication beginning in infancy with situations of joint attention and remaining important in adulthood. Among its other consequences, such communication, in effect, invites and enables a second person to share a stance toward something.

Drawing these observations together, there seems to be growing evidence to support the view that second-person relatedness is irreducible and important to human flourishing, that it is intimately associated with joint attention, and that there are strong candidates for at least some of the neural conditions and concomitants that enable joint attention in daily life. Do such discoveries have much bearing on ethical issues, however, such as the formation of virtues? Sharing a simple stance of attraction or repulsion toward a physical object or activity may indeed communicate to another person that the object is or is not good, but the objects of ethics tend to be abstract matters, such as justice or temperance. Can second-person relatedness also assist in forming a moral stance and hence cultivate virtue? This principle provides the proposed metaphoric understanding of Aquinas's account of theological *eudaimonia*, and therefore, the question is important to both philosophy and theology.

Researchers in social neuroscience often address pertinent issues using rather different terminology to that of virtue ethicists, but there are some experiments that may help to shed light on this question. Although such research is still new, there is

accumulating evidence that babies have a surprisingly sophisticated moral awareness, with 6- and 10-month-old babies at Yale University's Infant Cognition Center preferentially selecting "nice" over "naughty" puppets (Bloom 2013). What is significant about such findings is that babies inhabit a world in which their awareness of persons is presumably almost entirely second-personal, since it seems unlikely that they are able to carry out moral reasoning about persons in an abstract manner, and it has also been argued that they do not have properly formed first-person awareness until some time in their second year (Ferrari and Sternberg 1998; Keenan et al. 2003). So second-person relatedness is arguably already shaping moral awareness in the first year of human life. Supporting evidence for this conclusion comes from the difficulty of inculcating even basic virtues of self-preservation, such as the desire to eat in a temperate manner, in situations in which second-person relatedness is inhibited (Legge 2002).

Among adults, it is generally harder to disentangle the effects of second-personal relatedness from the conclusions of abstract moral reasoning. Nevertheless, there is some evidence that a moral stance is influenced, often in a sublimal manner, by the presence of a second person at the periphery of awareness, or even just a representation of a person such as a picture of a pair of eyes (Bateson et al. 2006). Moreover, even some choices that appear to involve only a single moral agent sometimes replicate patterns of second-person relatedness. As has been verified experimentally many times, "the act of breaking one's resolve can lead to a general dys-regulation of appetitive behaviour" (Wagner et al. 2011; cf. Herman and Mack 1975). These findings appear to be consistent the view that a personal moral resolution functions rather like a derivative of a covenant with a second person that, once broken, requires an act of reconciliation and a new resolution before a virtuous pattern of life can be resumed.

Conclusion and Future Directions

As the brief outline above has attempted to show, an account of moral theology developed by a Dominican friar in the thirteenth century has surprising relevance to contemporary work on the virtues and human flourishing, an approach that organizes many findings of social neuroscience around the principle of second-person relatedness. For a life shaped by a similar theological outlook today, this work may help to reinforce the value of specifically second-personal aspects of religious practices, such as the depictions of holy faces in sacred art, the use of music in shared liturgy, and the communication of religious teaching by stories of persons as well as propositions and rules. An analytically trained mind may be tempted to overlook the value of such practices, but they are not adventitious to a worldview in which theological *eudaimonia* is second-personal and culminates in divine friendship. Beyond its theological origins, however, there is hope that such research may help to contribute to a broader "Copernican Revolution" in the understanding of virtue ethics in the twenty-first century, "a shift in the locus of explanation from the first-person to the second-person perspective" (Pinsent 2012).

Cross-References

- ▶ Brain Research on Morality and Cognition
- ▶ Consciousness and Agency
- ▶ Impact of Brain Interventions on Personal Identity
- ▶ Mental Causation
- ▶ Moral Cognition: Introduction

References

Adolphs, R., & Janowski, V. (2011). Emotion recognition. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 252–264). New York: Oxford University Press.

Assal, G. (2001). Prosopagnosia. *Bulletin de l'Académie Nationale de Médecine, 185*(3), 525–535; discussion 535–536.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 2*(3), 412–414.

Belin, P. (2011). "Hearing voices": Neurocognition of the human voice. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 378–393). New York: Oxford University Press.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Malden/Oxford: Blackwell.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bloom, P. (2013). *Just babies: The origins of good and evil*. New York: Crown Publishing Group.

Bodamer, J. (1947). Die prosop-agnosie. *Archiv Für Psychiatrie Und Nervenkrankheiten, 179*, 6–53.

Campbell, J. (2005). Joint attention and common knowledge. In N. Eilan et al. (Eds.), *Joint attention: Communication and other minds* (pp. 287–297). New York: Oxford University Press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893–910.

Ciaramelli, E., Muccioli, M., Làdavas, E., & Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2*(2), 84–92.

Costall, A., & Leudar, I. (2004). Where is the 'theory' in theory of mind? *Theory & Psychology, 14*(5), 623–646.

Dalrymple, K. A., Corrow, S., Yonas, A., & Duchaine, B. (2012). Developmental prosopagnosia in childhood. *Cognitive Neuropsychology, 29*(5–6), 393–418.

Donald, M. (2001). *A mind so rare: The evolution of human consciousness*. New York/London: Norton.

Eilan, N. (2005). Joint attention, communication, and mind. In N. Eilan et al. (Eds.), *Joint attention: Communication and other minds* (pp. 1–33). New York: Oxford University Press.

Falk, D. (2009). *Finding our tongues: Mothers, infants and the origins of language*. New York: Basic Books.

Farah, M. J., Levinson, K. L., & Klein, K. L. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia, 33*(6), 661–674.

Ferrari, M. D., & Sternberg, R. J. (1998). *Self-awareness: Its nature and development*. New York: Guilford Press.

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *The European Journal of Neuroscience, 17*(8), 1703–1714.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives, 7*(2), 159–171.
Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience, 12*(9), 1187–1196.
Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), 593–609.
Gärdenfors, P. (2003). *How homo became sapiens: On the evolution of thinking*. Oxford/New York: Oxford University Press.
Heal, J. (2005). Joint attention and understanding the mind. In N. Eilan et al. (Eds.), *Joint attention: Communication and other minds* (pp. 34–44). New York: Oxford University Press.
Heilman, K. M., Leon, S. A., & Rosenbek, J. C. (2004). Affective aprosodia from a medial frontal stroke. *Brain and Language, 89*(3), 411–416.
Hein, G., & Singer, T. (2008). I feel how you feel but not always: The empathic brain and its modulation. *Current Opinion in Neurobiology, 18*(2), 153–158.
Herman, C. P., & Mack, D. (1975). Restrained and unrestrained eating. *Journal of Personality, 43*(4), 647–660.
Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience, 21*(7), 1229–1243.
Hobson, P. (2005). What puts jointness into joint attention? In N. Eilan et al. (Eds.), *Joint attention: Communication and other minds* (pp. 185–204). New York: Oxford University Press.
Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science, 9*(2), 131–134.
Hynes, C. A., Baird, A. A., & Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. *Neuropsychologia, 44*(3), 374–383.
Iacoboni, M. (1999). Cortical mechanisms of human imitation. *Science, 286*(5449), 2526–2528.
Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), e79.
Keenan, J. P., Gallup, G. G., & Falk, D. (2003). *The face in the mirror: The search for the origins of consciousness*. New York: Ecco.
Kirwan, M. L., White, L. K., & Fox, N. A. (2011). The emotion-attention interface: Neural, developmental, and clinical considerations. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 227–242). New York: Oxford University Press.
Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or a heuristic? *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society, 21*(6), 411–427.
Klin, A., Sparrow, S. S., Bildt, A., Cicchetti, D. V., Cohen, D. J., & Volkmar, F. R. (1999). A normed study of face recognition in autism and related disorders. *Journal of Autism and Developmental Disorders, 29*(6), 499–508.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911.
Kraut, R. (2012). Aristotle's ethics. In *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/win2012/entries/aristotle-ethics/. Accessed 19 August 2013
Kuhl, P. K. (2011). Social mechanisms in early language acquisition: Understanding integrated brain systems supporting language. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 649–667). New York: Oxford University Press.
Legge, B. (2002). *Can't eat, won't eat dietary difficulties and autistic spectrum disorders*. London/Philadelphia: Jessica Kingsley Publishers.
Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences, 8*, 529–566.
MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297–337.

MacIntyre, A. (2007). *After virtue: A study in moral theory* (3rd ed.). Notre Dame: University of Notre Dame Press.

McGilchrist, I. (2009). *The master and his emissary: The divided brain and the making of the western world*. New Haven/London: Yale University Press.

McInerny, R. M. (1993). *The question of Christian ethics*. Washington, DC: Catholic University of America Press.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology: CB, 17*(19), 1692–1696.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78.

Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33–35.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450.

Niedenthal, P. M., Eelen, J., & Maringer, M. (2011). Embodiment and social cognition. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 491–506). New York: Oxford University Press.

Nusbaum, H. C. (2011). Language and communication. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 668–679). New York: Oxford University Press.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, 91*(1), 176–180.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain), 223*(1232), 293–317.

Pinsent, A. (2012). *The second-person perspective in Aquinas's ethics: Virtues and gifts*. New York/Abingdon: Routledge.

Porter, J. (1992). The subversion of virtue: Acquired and infused virtues in the 'Summa theologiae'. *Annual of the Society of Christian Ethics, 12*, 19–41.

Reddy, V., & Morris, P. (2004). Participants don't need theories knowing minds in engagement. *Theory & Psychology, 14*(5), 647–665.

Roessler, J. (2005). Joint attention and the problem of other minds. In N. Eilan et al. (Eds.), *Joint attention: Communication and other minds* (pp. 230–259). New York: Oxford University Press.

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia, 45*(1), 124–143.

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia, 45*(13), 3054–3067.

Sperry, R. W. (1952). Neurology and the mind-brain problem. *American Scientist, 40*(2).

Stump, E. (2010). *Wandering in darkness: Narrative and the problem of suffering*. Oxford: Clarendon.

Stump, E. (2011). The non-Aristotelian character of Aquinas's ethics: Aquinas on the passions. *Faith and Philosophy, 28*(1), 29–43.

Tager-Flusberg, H. (Ed.). (1994). *Constraints on language acquisition: Studies of atypical children*. Hillsdale/Hove: Erlbaum.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9*(4), 483–484.

Tomasello, M. (1998). Reference: Intending that others jointly attend. *Pragmatics & Cognition, 6* (1–1), 229–243.

Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before speech: The beginning of interpersonal communication* (pp. 321–372). Cambridge/New York: Cambridge University Press.

Völlm, B. A., Taylor, A. N. W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J. F. W., & Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. *NeuroImage, 29*(1), 90–98.

Wagner, D. D., Demos, K. E., & Heatherton, T. F. (2011). Staying in control: The neural basis of self-regulation and its failure. In J. Decety & J. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 360–377). New York: Oxford University Press.

Werle, M. A., Murphy, T. B., & Budd, K. S. (1993). Treating chronic food refusal in young children: Home-based parent training. *Journal of Applied Behavior Analysis, 26*(4), 421–433.

Wimpory, D. C., Hobson, R. P., Williams, J. M. G., & Nash, S. (2000). Are infants with autism socially engaged? A study of recent retrospective parental reports. *Journal of Autism and Developmental Disorders, 30*, 525–536.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141–145.

Section XXI

Neuromarketing

Ethics of Neuromarketing: Introduction

102

Edward H. Spence

Contents

Abstract

This section examines ethical concerns that arise in neuromarketing. In particular, the entries in this section discuss ethical issues with regard to *brain privacy* (Steve Matthews) specifically, "the collection over time, and aggregation of private brain information, where the target loses control over its ownership and distribution" (Steve Matthews); the ethical concerns that Direct Consumer Advertising of Prescription Pharmaceuticals (DTCA) involving implicit persuasion through evaluative conditioning has "deleterious effects on autonomous agency" and that it has a negative impact on the "wider doctor-patient relationship" (Paul Biegler, Jeanette Kennett, Justin Oakley, and Patrick Vargas); and in addition, the Introduction of this section (Edward H. Spence) briefly examines how the association of brands with values raises ethical concerns relating generally to some of the ethical issues examined by the aforementioned authors in this section.

E.H. Spence
Centre for Applied Philosophy and Public Ethics (an Australian Research Council Special Research Centre), Charles Sturt University, Canberra, Australia

3TU Centre for Ethics and Technology, University of Twente, Enschede, The Netherlands
e-mail: espence@csu.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_101,

Assessing the Ethical Concerns of Neuromarketing

As Steve Matthews outlines in his chapter of this section, ▶ Chap. 103, "Neuromarketing: What Is It and Is It a Threat to Privacy?," peer-reviewed published articles on neuromarketing jumped from nothing in 2000 to 250 published articles and 150 neuromarketing companies in 2010 (Matthews, p. 9). Although there has been a sharp increase in neuromarketing publications and neuromarketing companies, Matthews argues that neuromarketing practices as such do not raise any special ethical worries or novel ethical concerns that were not previously present in product marketing and advertising.

For Matthews, the main ethical concern in neuromarketing relates to privacy and, in particular, brain privacy. According to Matthews "worries about brain privacy, seem, prima facie to be justified, but on closer analysis, fall away." He does, however, cautions that "a residual threat to privacy does remain: the collection over time, and aggregation of private brain information, where the target loses control over its ownership and distribution" (page 1628).

In my article "The Advertising of Happiness and the Branding of Values" (Spence 2013; Spence and Van Heekeren 2005) I am led to a similar conclusion with regard to the ethical concern raised by the implicit and explicit association of brands with values in advertising. Namely, that it is not the piecemeal association of brands with values that is the ethical problem but rather the widespread, systemic, and cumulative association of brands with values that is of ethical concern, that is, the concern that over time that association de-values and degrades aspirational social values such as *friendship* and *happiness* among others, by reducing them to the mercantile and consumer value of brands.

The central ethical problem Matthews focuses on in his chapter is that of brain privacy. Matthews' main claim for which he argues with great clarity and compelling arguments on the basis of the "principle of proportionality" (a philosophical principle also used efficaciously by Aristotle in his *Nicomachean Ethics*) is that "worries about brain privacy seem, prima facie, to be justified, but on closer analysis, fall away." He cautions, however, that "a residual threat to privacy does remain: the collection over time, and aggregation of private brain information, where the target loses control over its ownership and distribution."

Matthews also quite rightly draws attention to the effectiveness of certain communication strategies that operate at the interface of the internal brain environment of consumers (their brain states) and the external shopping environment in which they exercise their purchasing choices and make their purchasing decisions. As Matthews argues, empirical studies have shown that "Emotionally charged advertisements are more effective in generating memories of the brand. This makes plausible the view that neuromarketeers (*qua* marketers) are interested in brains-in-the-market; and the brain, hitherto inaccessible to market producers and advertisers, now takes its rightful place in the model to complete the picture."

Matthews' argument about the effectiveness of certain communication strategies used by neuromarketers and advertisers to generate emotionally charged advertisements that enhance memory retention and attention for brands, appears to dovetail

and lends empirical support to my own ethical concern regarding the association of brands with aspirational values. One such communication strategy, the association of brands with values, which I discuss in detail in my aforementioned article (Spence 2013), does seem prima facie at least to be effective in generating a positive emotional response or pro-attitude to particular brands through the association of those brands with particular aspirational values such as friendship, love, or happiness.

In their chapter of this section, Paul Biegler, Jeanette Kennett, Justin Oakley, and Patrick Vargas (for ease of reference Biegler et al. henceforth) focus on ethical problems that arise with regard to *Direct to Consumer Advertising of Prescription Pharmaceuticals (DTCA)*. According to them this is a controversial practice that is only permitted in the United States and New Zealand. The focus of their discussion is implicit persuasion through evaluative conditioning, which they argue has "deleterious effects on the autonomous agency that DTCA viewers bring to medicine choices, and on the wider doctor patient relationship" (page 1647). Their conclusion is that "There are good reasons to suppose that evaluative conditioning leads to more positive attitudes toward the advertised drug, and that such attitudes encompass inflated beliefs about drug safety and efficacy. Of grave concern, however, *is that the positive imagery deployed to produce such conditioned beliefs bears little substantive relationship with properties of the advertised drug* [emphasis added]. Given the materiality of these drug properties for people contemplating pharmacological treatment, such an unreliable influence on belief will likely undermine their justification, and antagonize the autonomy of resulting medical decisions." (Biegler et al., page 29)

The segment emphasized in italics in the above quoted passage is meant to draw attention to similar findings in my article (Spence 2013), albeit from a conceptual analytic perspective as compared to these authors' empirical findings. In that article, I raise similar concerns but specifically with regard to the association of brands and values. Such an association I argue is ethically problematic at the very least, because it is misleading and inaccurate at best and at worst false and deceptive as brands in themselves do not have the inherent characteristics of the aspirational values attributed to them by the advertising association to consumer products.

In their chapter, Biegler et al. concentrate their attention on implicit persuasion used in DTCA where individuals are unaware of the persuasive process being used. Although the persuasion is supraliminal and the resulting attitude achieved can be reported by the individuals subjected to this kind of persuasion technique, the mode of persuasion itself remains hidden (page 1649). They offer a nice example of this type of supraliminal persuasion from advertising. In advertising, they point out, "images of happy smiling people may be used to condition positive attitudes toward a novel product, for example, a new brand of shampoo" (page 1650). I would add that the same technique is used in the association of brands with values to which I have referred to earlier (Spence 2013). To the extent that I am right one would be justified in saying that the association of brands with values is a type of supraliminal persuasion where the resulting attitude can be reported by the individual but whose mode of persuasion remains hidden. Prima facie at least the empirical evidence

cited by Biegler et al. in support of their arguments for ethical concerns with regard to this type of implicit persuasion lends support to my own conceptual analytic findings in my article (Spence 2013) where I raise similar ethical concerns for the association of brands with values.

Based on various empirical studies which the authors examine in meticulous detail, Biegler et al. conclude that "There are strong reasons to hold that evaluative conditioning in DTCA, and other implicit persuasive techniques, comprise unreliable belief forming mechanisms that foster unjustified beliefs about pharmaceuticals."

According to the authors, this raises concerns about the overall effect that these techniques have on the autonomy of individuals subjected to these techniques, which may result in the "corruption of the deliberative process" in the choices they make (page 1654).

One final issue of great importance raised by the authors is skepticism about the role of rational agency itself in decision-making, which seems supported by various empirical studies cited by the authors, which show that, "notwithstanding our beliefs to the contrary, individuals really have little idea of why they make the choices they do" (page 1656). At first glance, this seems to undermine the concerns raised by the authors concerning the effect on autonomy and agency by the implicit persuasion techniques discussed in their chapter. However, although the authors accept the findings of the empirical results concerning the role of rational agency in decision-making, they quite rightly and perspicuously respond to such skepticism by pointing out that "evidence about the limits of rational agency and the need to counter irrational assessments of risk made by patients exposed to DTCA cannot constitute an argument for permitting implicit persuasion in DTCA. Rather it seems to constitute an argument against permitting DTCA at all" (page 1658).

Conclusion and Future Directions

Together the two chapters in this section raise measured but serious ethical concerns about the role and deleterious effects of neuromarketing on brain privacy, as well as concerns about the role of implicit persuasion through evaluative conditioning as used in Direct Consumer Advertising Pharmaceuticals (DTCA) and its related negative impact on the wider doctor-patient relationship. In addition, within the context of those concerns, I have raised related ethical issues concerning the general problem concerning the systemic and widespread advertising practice of associating brands with values, which I have argued can be seen as a general implicit persuasion technique through evaluative conditioning. Together these articles provide novel research that raise ethical concerns about neuromarketing techniques that should be taken seriously. Further empirical research to evaluate how the association of brands with values degrades those values would be useful in assessing the impact it has on our social perception and appreciation of values. As well, it would also be very useful to relate that empirical research on brands and values to the type of empirical research on neuromarketing referred to in the chapters of the authors in this section.

Cross-References

- ► Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
- ► Ethics of Neuromarketing: Introduction
- ► Ethics of Implicit Persuasion in Pharmaceutical Advertising
- ► Human Brain Research and Ethics
- ► Neuromarketing: What Is It and Is It a Threat to Privacy?

References

Spence, E. (2013). The advertising of happiness and the branding of values. In M. Boylan (Ed.), *Business ethics* (2nd ed.). Upper Saddle River: Pearson/Prentice Hall.

Spence, E. (lead author), & Van Heekeren, B. (2005). *Advertising ethics*. Upper Saddle River: Pearson/Prentice Hall.

Neuromarketing: What Is It and Is It a Threat to Privacy? 103

Steve Matthews

Contents

Abstract

This entry has two general aims. The first is to profile the practices of neuromarketing (both current and hypothetical), and the second is to identify what is ethically troubling about these practices. It will be claimed that neuromarketing does not really present novel *ethical* challenges and that marketers are simply continuing to do what they have always done, only now they have at their disposal the tools of neuroscience which they have duly recruited. What will be presupposed is a principle of proportionality: marketing practices are morally objectionable commensurate with the degree to which they impugn the moral sovereignty of market actors. With this principle in mind, it is important to consider the literature which is skeptical about the potential for

S. Matthews
Plunkett Centre for Ethics (St Vincent's and Mater Health Sydney), Department of Philosophy, Australian Catholic University, Sydney, NSW, Australia
e-mail: Stephen.Matthews@acu.edu.au

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_154,

neuromarketing to be successful. If its claims are overblown, as will be suggested, then the ethical threat neuromarketing is said to pose can be viewed also as overblown. An area that has worried many is that neuromarketing poses a threat to brain privacy, and so an analysis will be given of the nature of this threat, given the principle of proportionality. It will be argued that worries about brain privacy seem, prima facie, to be justified, but on closer analysis fall away. However, a residual threat to privacy does remain: the collection over time, and aggregation of private brain information, where the target loses control over its ownership and distribution.

Introduction

The literature on neuromarketing variously describes it as the "...application of neuroimaging techniques to sell products" (Lee et al. 2007, p. 200), the use of neuroscience to gain "...powerful insights into the human brain's responses to marketing stimuli" (Murphy et al. 2008, p. 293), "...the application of the findings from consumer neuroscience within the scope of managerial practice." (Hubert and Kenning 2008, p. 274), and "...applying the methods of the neurology lab to the questions of the advertising world." (Thompson 2003, p. 53). Although somewhat helpful in landing us in the right territory, these definitions are loose and sufficiently non-coextensive for neuroethicists to be rightfully dissatisfied. (To be fair, each of the writers above sought to identify their subject matter in quick and dirty terms). Part of the motivation here is to become clearer about the nature of neuromarketing in order to assess that practice from an ethical standpoint; a set of defining conditions for neuromarketing will not be proffered, but rather, the intention will be to provide some description of the institutions and practices that are both presently associated with neuromarketing, as well as those that seem a reasonable possibility in the near future, so as to evaluate their ethical significance.

Definitions

The term "marketing" refers to those processes or activities that facilitate the movement of a product or service by focusing on its value in an exchange setting. In that broad sense, it refers to the design and production of goods and services, their distribution and pricing, and the communication strategy set forth in the market. Thus construed, advertising (the communication strategy) forms a subbranch of marketing. Advertisers communicate a message about a product or service ultimately to increase sales or market share. A marketer need not be narrowly focused on the persuasive aspect of a business. For instance, a marketing department of a company typically works in close proximity with company operations concerned with product design or content (Kotler and Keller 2012; see part 5). Currently, neuromarketers are less able to retrieve neuroscientific data in natural settings and tend to test subjects in the lab, typically for responses to such things as brand choice, pricing, product design, the

buying environment, responses to different kinds of advertisements, what grabs the consumer's attention, etc. This is neuromarketing devoted to *understanding* the brain-mind of the consumer, in contrast to a practice in which they might seek more directly to influence the consumer. The intention of neuromarketers to develop this capacity (securing brain response data in the market, in real time) raises an important strategic question for them: at which point, along the chain from design to use of a product, should they extract this data? (The place they seem most interested in, as discussed below, is point of sale). The ethical questions here will be around the invasiveness, and transparency, of this strategy.

Neuromarketers, then, although almost wholly devoted to recruiting neuroscience techniques to identify features of the consuming mind for market advantage, may also in the future engage in the business of persuading and influencing directly. Neuromarketers who aim to influence (neuroadvertisers let's say) would be more at risk for ethical breaches because they are engaged in advertising behavior that is inherently morally risky. As described below, this consists in boundary crossings, nondisclosure, and targeting of customers. For neuroscience that seeks only to identify features of consumer cognition, these risks subside. Neuroscientists here conceive of themselves as generating knowledge about the mind of the consumer, and marketers may, and do, latch onto this research. These neuroscientists are of course not engaged in marketing, and the distinction between research and this kind of engagement is clear enough by observing the place from which they work: the academy. (Things are less clear when qualified neuroscientists are employed by neuromarketing companies or form independent private consultancies themselves). Almost all neuroscientists who work in the academy and who research marketing questions do so not to the exclusion of other related interests; they would be dismayed were they referred to as "neuromarketers." A neuromarketer, in my usage, is a person with an interest in applying neuroscience ultimately to derive a commercial outcome. A neuroadvertiser is a kind of neuromarketer, one whose role is less remote from the shop floor than a neuromarketer who is not a neuroadvertiser.

(It should be recognized that these distinctions are impure: there may be cases where an individual occupies a dual role, as marketer and as academic, or whose role is morally ambiguous (say when the research is funded by a private corporation). These cases raise a familiar ethical question for such a stakeholder: are commercial interests in this research contaminating its intellectual purity? Notwithstanding its importance this question will not be pursued here).

Neuroadvertising: A Hypothetical Example

Consider an elaboration of a hypothetical example of neuroadvertising (Wilson et al. 2008). Although most of the buzz in neuromarketing has been, and still is, around fMRI scanning, other technologies may be used, including Quantitative electroencephalography (QEEG) and Magnetoencephalography (MEG). QEEG, say its advocates, "...is simpler and less expensive to use and enables recordings to be made in a wide range of natural environments" (Lewis and Brigder 2005, p. 37).

Nevertheless, QEEG is still somewhat cumbersome. So let us imagine the currently impossible situation of a technology that neuroscreens subjects remotely and potentially secretly. Marketers may, then, envision the use of such a thing in something called the "intervention" phase of a persuasion model of neuromarketing (Wilson et al. 2008, p. 395). In designing a large department store, a marketer might imagine the impact of the stimuli on the inchoate consuming brain as it enters the store, motivated as the marketer is to maximize the amount of money to be shifted from the target consumer's bank account to the storeowner. The marketer must, therefore, understand the impact of these stimuli and design accordingly. To effect this, say Wilson et al. (2008, p. 398), "...retailers may neuroscreen potential customers upon entering, registering reactions to what they see, hear, feel, touch, taste, and/or smell, and combining these measurements and outcomes with previous readings based on earlier visits."

In this example, the marketing aim is to use neuroimaging remotely and directly as a tool to smooth the progress of the transaction in real time and in real commercial space. What is ethically bothersome about such a practice? The all-important distinction is of course between the case where full disclosure takes place and cases where less and less disclosure is evident. In the case in which no disclosure takes place, three elements of ethical relevance stand out. First, this is a clear case in which the seller is manipulating the environment to sell goods. Is that ethically objectionable? Is it sui generis? The answers, surely, are "no, not really" and "no, it is standard practice – it is well known as part of the institution that constitutes market practice...buyer beware!" The second is that this is a case where the consumer's lack of information about the retail environment prevents her from consenting to aspects of the transaction there. This is more serious because the seller is now operating with an unfair advantage given that the buyer's natural competence to operate in that situation has been undermined without his permission.

A third element is more serious yet. An ethically all-important feature of the case is that the real-time brain snapshots are compared with previous readings that were made of that same shopper. Don't forget that this is a case in which no disclosure takes place, and so it is a case where this shopper's neuro-profile is being recorded, stored, manipulated, and compared, without his permission and quite possibly against his interests. This is, then, a prima facie breach of privacy (of both one's person and one's information). An important argument against neuromarketing, then, is this: at its worst it will further erode an ethical boundary between self and world at a point that was formerly thought inviolable: the skin (Fischbach and Mindes 2011, p. 361). Suppose some shopper, standing quite still, and with a deadpan facial expression, views a pale blue dress. From her behavior, then, there are no real grounds to infer her thoughts. The scanner, meanwhile, revealing high activity in the orbitofrontal cortex, anterior temporal pole, and amygdala, tells a story about visual memory and emotion. This pale blue dress has evoked an emotional response in the shopper, likely involving a memory of some past episode.

Now there are several difficulties with what has just been described. First, the degree of specificity that attaches to snapshots at that level of description – "visual memory," "emotion" – is far too general to generate meaningful assumptions about

what, *exactly*, this person is thinking. That is problematic for a marketer working out how to capitalize on such impoverished information, and it is, insofar as that is true, thereby a less worrisome state of affairs for the ethical evaluation. Intuitively, do you object more to the fact that another has determined your emotional state from secretly scanning your brain or to (say) knowledge that some shop assistant, who knows you as a regular, was peering at your sad face? There does not seem to be a great deal of difference. But, some may, intuitively, disagree.

So, let's consider some data. In a 1997 study (Nwahukwu et al. 1997), an examination was made of people's evaluations of advertising strategies. The significant determinants were given as consumer sovereignty and autonomy and the harmfulness of a product. On the question of autonomy, emphasis was on the ability one has to recognize and neutralize the manipulations that are taking place. Yet, just as no manipulation is taking place while someone secretly peers at my sad face, no manipulation takes place as the scanner snaps my glowing amygdala. No manipulation means there is nothing to neutralize. Now one might respond here that given a certain level of brain-scanning accuracy, there is, at least, *potential* manipulation, say in the case where the machine provides much more detail, and a person in a monitoring booth retrieves this detail and passes it on to a salesperson who then says "bring back memories does it?" Suppose the targeted shopper asked the sales assistant "yes, reminiscing. . .how did you guess?" and was told "you just looked lost in the past" or some such thing. This kind of deceit would naturally send a shiver up many a Kantian spine. And those less exercised by the deceitfulness of the manipulation would perhaps also worry that in lowering the standards of transparency, the stability of the institution of market transactions might take the first slip down a worrisome slope.

Novel Ethical Problems?

Take a step back. This example is fictitious and designed to explore an ethical implication. In particular the unlikely presupposition was put in place that scanning takes place secretly. Nevertheless the case is instructive because – based on a principle of proportionality – the degree to which the practice is ethically dubious can be seen, roughly, to track the successfulness of the manipulation. Now, is it worrying that surreptitious scanning of brains will enable manipulation of customers? This question has two interpretations. One is that scanning raises an ethical difficulty that is novel. The other is that it exacerbates (perhaps significantly) an existing ethical dimension to the marketing practice. There are good reasons to think that neuroscanning does not raise a novel ethical challenge and that the degree of exacerbation depends entirely on the type and degree of success of the scans, as just discussed.

Neuromarketers are attempting to get access to information about the situated cognitive profile of consumers in order to understand the conditions that best motivate purchasing behavior. How new is this? Marketers have been attempting this *kind* of thing for a very long time indeed. Evidence of advertising, in one form or another, suggests the practice may have existed in ancient Egypt, Greece, and

Rome, and perhaps even earlier (Bhatia 2000). Modern advertising techniques are dated from the late 1800s. The most modern techniques in which members of the academy (usually those from the social and psychological sciences) join with marketers to inform them about what motivates the consumer go back quite a stretch to a period immediately after the Second World War. The remarkable book by Vance Packard called "The Hidden Persuaders," from 1957, discusses in great detail a range of practices occurring even then. It describes a "depth approach" to marketing in which insights about the subconscious nature of purchasing decisions are gleaned. In the depth approach it was recognized that many buying decisions are generated by processes that bypass the rational self. Insofar as that is true, campaigns were designed to cause purchasing behavior in whatever way worked, rather than cause purchasing behavior via the conscious reasoning mind. Thus, the real reason for deployment of a persuasive technique was worked out in advance by scheming marketers (p. 31). Packard continues:

> Certain of the probers. . .are systematically feeling out our hidden weaknesses and frailties in the hope that they can more efficiently influence our behaviour. . .[some are] probing sample humans in an attempt to find how to identify, and beam messages to, people of high anxiety, body consciousness, hostility, passiveness, and so on. A Chicago advertising agency has been studying the housewife's menstrual cycle and its psychological concomitants in order to find the appeals that will be more effective in selling her certain products. . . The same Chicago Ad agency has used psychiatric probing techniques on little girls. Public-relations experts are advising churchmen how they can become more effective manipulators of their congregations. In some cases these persuaders even choose our friends for us. . . (pp. 32–33)

For those who imagined that the alliance between marketers and some members of the academy was a relatively modern incarnation dating back only two decades or so, Packard's account may come as quite a surprise. He painstakingly details the practices of an already highly sophisticated and institutionalized approach to developing ways of bypassing the consumer's conscious awareness of what is taking place within a commercial transaction. Suffice to say, neuromarketing is simply an expected recruitment of an available technology that widens an already impressive suite of existing techniques for hidden persuasion. The aim is the same. The means is new. It might be thought, then, that the only question to ask, of ethical relevance, is one about how successfully neuromarketing fulfills the ever-present aim of probing the vulnerable consumer. That is a key question, no doubt, but as mentioned, another important area raised is brain privacy. There is good reason to think that worries about brain privacy are, to a large extent, overblown.

In the present context, some morally dubious advertising practices include those that lead to boundary crossings, targeting of consumers, and nondisclosure (Kennett and Matthews 2008). What exactly are these, and what is their relevance to neuromarketing? Advertising messages may be restricted to conventional spaces such as billboards or breaks between television programs, or they may intrude into noncommercial areas, as happens when product references are inserted into movies, novels, or creative works. People objecting to the latter are worried about boundary crossings. Undisclosed advertising examples include ads disguised as public

announcements; or comment, but paid for by a company; social actors (people in bars, perhaps even friends) secretly buzzing product; or subliminal messaging. Targeting occurs when an audience is selected because their profile makes them more likely to be receptive to the message. Companies know that advertising is less cost-effective when more of the targets ignore the message which occurs in mass campaigns. Most advertising campaigns have a degree of targeting – TV ads for toys don't run at midnight – and that's plain good sense. It is targeting of the vulnerable that is problematic. In the ethical evaluation of neuroadvertising, it is important to focus on questions of brain privacy where boundary crossing, targeting, and nondisclosure all feature.

Paradigm Cases

In neuromarketing (more broadly now), the focus widens to consider such things as product and service design, distribution, and pricing and their impact on consumer evaluations of risk and reward, their choice making, and the mind-brain processes leading to purchase decisions. Another question is whether scanning could be applied to other stakeholders in a market, for instance, employees. It should be noted that if scanning is regarded as a form of lie detection, legal hurdles will prevent this. In the USA, for instance, the Federal Employee Polygraph Protection Act 1988 prohibits employers from requiring or suggesting such a test be taken. And Murphy and Greely (2011, p. 650) discuss a recent district court ruling in the USA (United States v. Semrau 2010) that excluded fMRI-based lie detection material because of a failure to meet the standards of admission of scientifically based evidence set down federally.

A study in 2004 is regarded almost as archetypal in the field of neuromarketing. Researchers observed preferences for one brand of soft drink over another, where (in condition 2) the only relevant variable was brand knowledge of one of the samples, and where previously (in condition 1) this variable was absent. In condition 2, subjects displayed a statistically significant preference for the known-brand samples. In condition 1, fMRI imaging revealed consistent activation of the ventromedial prefrontal cortex (an area associated with decision-making and risk assessment), but in the brand-cued condition, in addition to the VFC, other areas of the brain were recruited – the hippocampus (involving processing of, inter alia, episodic memory information), dorsolateral prefrontal cortex (associated with, inter alia, affectively laden social judgments), and midbrain (associated with, inter alia, habits and motivation). The researchers suggested that two systems operate in concert to potentially bias decisions based on cultural meanings. When a brand is rich in such meanings, its associations work their way into memory content. These are affect-laden information pathways which have the potential to bias preference judgments, said the researchers (p. 385), and it is known that the recruited brain areas are implicated in cognitive functions of this type (McClure et al. 2004). The study is regarded as significant because it provided hard evidence – *brain* evidence! – that a particular kind of advertising that establishes a certain kind of

brand reputation is internalized and processed by the consumer in decision-making over product. So, it provided marketers with reassurance that the billions spent on brand promotion and loyalty campaigns was not only money not wasted but indeed important territory to battle it out for the competitive edge.

Neuroscience is thought to be able to help in other areas of the market, for example, in pricing policy. In lay terms the negative price effect refers to the tendency for demand of a good to reduce as its price increases, and the positive price effect refers to an increased demand. The psychology of this seems straightforward: on the one hand higher prices represent a net loss, and demand reduces; on the other hand higher prices signal higher quality and prestige offsetting loss perception and a rise in demand. These effects are, therefore, category sensitive and dependent on consumer knowledge of the market. So, for instance, positive price effects often obtain for luxury consumables such as wine or cars. Negative price effects obtain most clearly for familiar everyday substitutable items of known and stable quality, like a range of foodstuffs, such as milk, eggs, or biscuits. Since it is not always clear which effect will obtain in the pricing of certain types of goods, marketers are keen to have the best neuro-motivational analysis of potential consumers in order to get their pricing approach set to maximize profits.

A well-known study in this connection comes from Plassmann et al. (2008), who were interested in the neural correlates of the positive price effect. (See Knutson et al. (2007) for an fMRI study of the negative price effect). In brief, subjects' appraisal of wine quality was sensitive to price as the positive price effect might predict. Taste sensation experiences, unlike the uninterpreted raw feel itself, are complex. For taste sensation experiences, there are three separable neural systems: what takes place at the cellular level of taste buds, the neural pathways that carry the elicited transmitter, and those primary brain regions into which these transmitters project, namely, the brain stem, limbic areas, and some higher cortical regions. In the Plassmann study subjects who experienced the expensive wines as being of higher quality displayed cortical recruitment in their fMRI scans.

An Unstable Situation

Neuromarketers find these kinds of results to be of enormous interest because they seem to show evidence that market evaluations are modulated by brain regions thought responsible for higher cognitive function. More specifically, these results confirm, for marketers, that buying decisions can be biased by cognitive processes that have little to do with the internal composition of the goods themselves (given that a certain threshold of quality is achieved) and a lot to do with exogenous features of the purchasing event.

A number of pragmatic questions are here raised. For instance, why is it that evidence from neuroscience is deemed especially useful and reliable for application in marketing practice? Why are marketers so focused on an internal feature of commerce, the consumer brain? And why do they think this information is actually and potentially useful?

Neuroscience explanations are alluring. Perhaps they are because of the hope of mind reading – a capacity science (allegedly) might have to identify and describe a person's thoughts, in particular her beliefs and desires. Another reason might be that our brains don't lie. Agents may not really know why they act, and they may not always have the capacity to explain why they acted, and that is because in various contexts the potentially dissociable subpersonal mechanisms implicated in behavior are revealed for just that, and the connection between (articulable) reasons and actions is lost. For instance, a study by North et al. (1999), showed that participants' buying decisions – between bottles of German and French wines that were otherwise the same – were significantly affected by background (German or French) music, yet only one of the forty-four participants said so, and 86 % explicitly denied that music had anything to do with their choice. So, although agents sometimes cannot nominate the real causes of their actions, neuromarketers continue to be dazzled by the apparent fact that big machines can. This continues to be an enormous attraction, but there is reason to be suspicious.

The attractiveness of neuroscience in the market can be measured. Typing "neuromarketing" into Amazon Books yielded 219 hits (2/11/13). These are, in general, not reprints, and nearly all have publication dates from around 2009. It doesn't take long to find examples of hype, especially online. Consider this gem from CNNMoney: "And neuromarketing scans are very good at pinpointing exact points of reaction. Pradeep cited a study for an unnamed company that makes chips and salsa. They found that the moment a snacker lifts a salsa-covered chip to his mouth, before taking a bite, 'that moment is extremely evocative for the brain. Your brain just goes nuts.'" In 2000 there were no peer-reviewed articles, no Google hits, and no neuromarketing companies. In 2004 there were 5,000 Google hits, climbing almost vertically up the y-axis. In 2010 there were 250 published articles and 150 neuromarketing companies (Plassmann et al. 2012). The suspicions about the overvaluation of neuroscience explanations can be investigated as well. A paper by Weisberg et al. in 2008 showed that students and nonexperts tended to find flawed neuroscience explanations of psychological phenomena more satisfying than explanations lacking these flaws.

Summing up, this is an unstable situation: neuromarketing is nascent, growing exponentially, and, by the marketers, seemingly highly trusted. However, many think there is a reason to hesitate and reflect on what neuroscience can really do here; others are outright skeptical, and others think there is something fundamentally ethically misplaced about applying neuroscience in a commercial context. For example, Fischbach and Mindes (2011, p. 361) write that neuromarketing "is troubling" because it involves "distortion and potentially inappropriate commercialisation of science." On the question of skepticism, most of the focus is not on the accuracy, or design, of the science itself, but rather the estimation of the value of the information it provides for the market. And here, the thought is that, just like Weisberg's subjects, market nonexperts' bedazzlement is preventing clear thinking about what neuroscience is able to deliver *in principle*. Paul Wolpe (Director of the Emory Centre for Ethics), for instance, has commented in a New York Times interview that neuromarketing is really only a kind of "pop" neurology for product

positioning, and it will not get people to buy or buy more. One also gets the feeling that much of the overblown confidence about neuromarketing's possibilities issues from the market and the more restrained estimations issue from the laboratory – a kind of institutionalized version of both the endowment effect and the Dunning–Kruger effect working together. On the endowment effect (roughly, where people overvalue what they own), it seems likely that neuromarketing companies' financial and psychological investment in the practice bolsters their willingness to accept its results. On the Dunning–Kruger effect, according to it, poor performers in particular grossly overestimate their abilities and intellect precisely because their incompetence deprives them of the acumen needed to identify their own weaknesses (Ehrlinger et al. 2008).

Still, to be fair, taking seriously the nascent state of the science is important, and even claims about "in principle," possibilities have to be made from behind a firewall of ignorance. Compare the failures in the early days of computing to appreciate Moore's Law in relation to the biannual doubling of integrated circuits (Moore 1965). Moreover, consider the limitations of mechanical emulations of computing functions relative to electrically based switching components, the latter allowing novel functionality based not on a principled distinction between one kind of computing and another, but a simple difference in concrete realizability. So, there are lessons here for the overly pessimistic as well.

In addition, drawing on an earlier distinction between persuading the consumer and understanding what motivates the consumer is instructive. To argue skeptically that neuromarketers will never locate the so-called buy button constitutes a straw man argument when neuromarketers are attempting no such thing in the first instance. And many neuromarketers are keen to emphasize just this important distinction. According to two neuromarketing consultants,

> The use of brain-imaging will never enable marketing professionals to discover that Holy Grail of market research, a 'buy button' – some mythical region of the brain which need only be stimulated to compel consumers to purchase a product whether or not they actually want to do so! It will never be found because, of course, it does not exist! More realistically, we believe, Neuromarketing offers the prospect of gaining a better understanding of how the brain responds in a wide variety of everyday situations. In addition to proving of great commercial value such research offers the possibility of increasing our knowledge of brain function among a non-clinical population as it extends powerful medical technologies into a new and challenging area of research.

The quote is from David Lewis and Darren Bridger (2005, p. 37), who seem careful in their assessments of the limits of neuromarketing, though at the end that circumspection seems to dissolve into a flourish designed to bolster confidence in what they claim neuromarketing can achieve.

Skepticism

One worry that has been expressed is that neuromarketing doesn't tell us much we didn't already know or at least strongly suspect on independent grounds (see, e.g., Levy 2009, p. 11). Erk et al. (2002) used fMRI to study the brain activation patterns

of 12 male subjects who viewed photos of different classes of cars, then rated them for attractiveness. They had hypothesized that sports cars are associated with wealth and social dominance, and since dopaminergic reward circuitry is implicated in social relations like dominance and social status, they said, it would not be surprising if brain correlates were found associating sports car viewing with the relevant circuitry. This is indeed what they found. In particular they found that the level of perceived attractiveness of a car was correlated with the level of activation in the ventral striatum, a key brain site from which dopamine projections move into other regions implicated in motivation. These findings are important, so it is said, because of an inference linking product design to reward to purchasing behavior. Unfortunately, this is an expensive way to find out something discoverable by asking a car dealer. Less facetiously, Erk's hypothesis could be tested by social psychologists collecting the relevant data on purchasing patterns, earning capacity, and psychological profiling.

Now this is not to say that there is *no* reason for engaging in studies of the kind Erk et al. performed, in case they are viewed legitimately as baby steps for more sophisticated ones later. Paradigms in normal science take time to build and develop, and so many early studies serve as foundations for them and for further work incorporating these earlier accomplishments.

Levy (2009) raises a novel objection to the uses of neuromarketing. Stepping back, ask first: what is at the core of the ethical objections to marketing practices? The answer is that they are objectionable when they bypass autonomous decision-making in the marketplace. If agents are manipulated into buying a product, they have not consented to it. If that is at the core, and surely it is, then the ethical test for market practices at the buyer-seller interface is whether manipulation is present. Levy says that insofar as we are overly focused on the internal interventions of neuromarketers, we risk overlooking the most powerful manipulative techniques marketers deploy in the *external* environment.

There is a type of consequentialist principle operating here, namely, to pay close attention to the weakest empirical point at which there may be ethical breaches. If one's attention fastens on to the potential for neuro-manipulations, one may well lose sight of those weaker external points. Arguments in applied ethics must identify the relevant moral principle, and they must determine the empirical conditions in which it is likely to be compromised; less obviously, arguers must remain sensitive to an epistemic bias concerning the most likely source of their ethical concerns. In the present context a sensible policy to deal with the last point would be that observance of the distinction between agent and environment makes trouble when evaluating the ethics of neuromarketing. Levy goes on to make a connection between Baumeister's ego depletion hypothesis (1998) – willpower is a kind of exhaustible resource – and manipulations in the shopping environment that, albeit implicitly, exploit the vulnerable consumer. This thesis has also been connected with self-regulation and impulse buying (Vohs and Faber 2003).

In this connection it is somewhat surprising that little, perhaps no, explicit link is made in the literature between addiction neuroscience and neuromarketing. This is particularly so when one considers the legal availability of addictive products such

as alcohol. One might have thought that neuromarketers would be interested in the neuropsychology of consumers who develop a habit, one in which consuming the product of choice occurs as a kind of ritualized sequence or action repertoire. They might be tremendously interested in the state of a brain that results from that kind of life. David Nutt (2012, p. 136) has recently suggested an elegant model in addiction neuroscience in which pull factors, higher-than-expected pleasurable reward (dopamine), reduced suffering (endorphins), memories (GABA/Glutamate), and meanings (serotonin), combine with push factors, impulses (noradrenaline), compulsions, and withdrawal (multiple neurotransmitters), which then all feed into a state of wanting. The identification of the addicted brain is the identification – by many marketing standards – of the ideal consumer. Notwithstanding the legality of the products neuromarketers are concerned with, researching the neurological conditions of addiction for a commercial outcome is presumably not seen as ethically defensible. This would seem, on the other hand, to represent what has been called a dual use dilemma for addiction neuroscience. But that thought cannot be pursued here (see Miller and Selgelid 2009).

Are There Grounds for Moral Disquiet?

Bringing together the threads of the discussion, so far, there are indeed grounds for *some* disquiet in regard to the future of neuromarketing. What might be potentially morally problematic here? Assume, first, surely not implausibly, that neuromarketers, *qua* marketers, are designing a model to maximize the market share obtainable by business X, where the consumer (including his brain) and aspects of the consumer environment are conceptualized as parts internal to the same system. Then, they are not really interested in agents separate from the shopping environment; they are interested in oiling the components of the system to optimize the profits of X. So, they will be interested in the work of Baumeister, and they will be interested in the work on subliminal brand priming (Murawski et al. 2012; Winkielman et al. 2005), work done on the effects of the hormone oxytocin on trust (Zak et al. 2004, 2005), and any other work that might facilitate the movement of a product or service in an exchange setting. They will of course be especially interested in experiments that utilize brain scans, for example, as a way of confirming the effectiveness of certain communication strategies. (As did Ambler and Burne (1999) and Ambler et al. (2000) when testing the relevant role of cognitive versus affective strategies in brand advertisements by administering beta-blocker drugs to participants (which dampen affective responses) to show that in the normal case emotionally charged advertisements are more effective in generating memories of the brand.) This makes plausible the view that neuromarketers (*qua* marketers) are interested in brains-in-the-market; the brain, hitherto inaccessible to market producers and advertisers, now takes its rightful place in the model to complete the picture.

Assume also the distinction between brain types and tokens. When McClure et al. did their research, they sampled many individuals because they were

interested in truths about what the human brain does *in general*, in response to like-stimuli. The thought here is that such general truths carry more robust predictive value and so more value to the marketer looking for efficient advertising strategies. That is correct enough, but the level of generality here is insufficient to properly and effectively target the individual customer. Even the environmental manipulations have their limitations. Absent specific historical information about particular consumers who present within a store, a store that has maximized the product-moving effectiveness of its design (e.g., in the placement of its products), will nevertheless reach saturation level with respect to the possibility of further sales. If the store had access to information about particular regular customers – their brain scans and buying patterns, all crossmatched – more targeted and effective marketing strategies could no doubt be achieved. If the earlier scenario involving real-time mobile scanners were in place, it seems not unlikely that, for instance, a shopper's emotional response could be discerned with reasonable accuracy, which might be seen as useful information for a shop assistant. This would be brain level, token-specific, historically accurate, up-to-date data that delivers a profile of the consumer. Such information may be kept on a database, triggered in real time via face-recognition software, and delivered at the right moment and setting to calculate the best sales approach given the current state of the consumer.

That description seems to pinpoint the empirical conditions that most worry those who are concerned about a consumer's right to control his or her privacy, both informational and so-called perceptual privacy (Matthews 2008, p. 130). However, this worry can be downplayed by analyzing away the grounds of the fears by (1) talking about what is really at stake in questions of privacy, (2) thinking about what "mind" refers to in "mindreading," and (3) noting again the inherent deficiencies of brain imaging for the alleged neuromarketing purposes. Should this attempted debunking allay the worries of Fischbach and Mindes (2011, p. 361) and others that neuromarketing could lead to an "invasion of brain privacy"? No, not completely. It is important to sound a note of caution about the possibility that companies will gain control of neuromarketed personal records built up over a period of time, leading to aggregated datasets of quite detailed information. In this form it may be stored for use, distributed, traded, sold, and as sometimes happens misused, accidentally leaked, and so on.

(Questions of brain privacy also arise in neurolaw, and sometimes there is a reversal from protection of brain information to attempts at having the contents of one's brain utilized for, in some cases, legal exoneration. In Harrington vs Iowa, testimony making use of "brain fingerprinting" was introduced in an attempt to exonerate a convicted murderer. Lawyers making use of the so-called P300 effect attempted to show, through evidence given by Dr Lawrence Farwell, that the defendant did not have knowledge of the murder. The P300 effect involves the emission of a signature EEG-recognizable pattern (coming 300 milliseconds poststimulus), and (controversially) it is claimed to reveal information stored in a subject's brain (see Peters 2005).

Skepticism Again

First, consider (2) and (3) above, taken together. The literature in relation to the accuracy and usefulness of brain imaging data varies. In Canli (2006), neuroimaging, he claims, predicts some behavior with more accuracy than self-report. Others, such as Gazzaniga (2006, Chap. 7) and Levy (2007, Chap. 4) are not so confident. The question of accuracy is partly open and partly a conceptual matter. On the empirical openness point, the accuracy obviously depends on technological improvements as discussed above. However, consider a known feature of fMRI that would seem an impediment to neuromarketing utility. They deliver relatively high-value spatial resolution scans, though getting the correct sample frequency (temporal resolution) is problematic given that blood takes a relatively long time to redistribute within the brain, and so this "hemodynamic lag" negatively impacts the capacity to make accurate brain-mind correspondences where the possession of temporal synchrony is salient. Moreover, since their predictive error is a function of the signal change between two conditions, the sample size within an experiment is technologically critical to the statistical power assigned to the result, though it depends on the relevant cognitive task (for the technical details of all this, see Desmond and Glover 2002). Suffice to say, these observations suggest a limitation for fMRI-type scanning as a means to identify "one-off" buying situations involving single individuals operating in novel conditions. (And all of this is not to even mention that they are car-sized, non-portable, dangerous, and immensely expensive machines (between one and three million dollars)). There are other difficulties regarding the limitations of fMRI, including extrapolation of correlates to different population types, different circumstances, and sampling biases arising from an inability to compare published results with unpublished null results from similar experiments (see Farah et al. 2008).

On the conceptual question, what is supposedly being read when our shopper enters the store? A useful first distinction here is that between dispositions and states. The store brain scanner might detect activation in posterior regions of the hypothalamus, accurately indicating a state of thirst, and that *could* be useful information. But the scanner won't be so good at predicting a range of specific behaviors based on descriptions of static states, even very detailed ones, of my brain anatomy. In other words it won't be very good at identifying how I am *disposed* to act. This is partly to do with the fact that it would not take into account the range of environmental triggers of such dispositions but also because there are too many possibilities. One way of unpacking the notion of a disposition (for present purposes anyway) is in terms of a set of counterfactual propositions true of some agent. The question, then, might be whether the scanner (in contrast with cheaper garden variety methods involving say talking to people) could generate data for deciding on the truth of these:

Were this shopper to X, he would Y
Were this shopper to be X'd, he would Y

And so on. On the question of states, even here there are difficulties. If, for instance, moods are regarded as internally generated background affective states, which, of conceptual necessity, last longer than certain emotions, then it is doubtful

scanners will accurately reveal mood types in normal subjects. And it would seem impossible to identify the intensity of a mood relative to what is normal for that individual. The scanner will be better at identifying certain kinds of emotion – for example, disgust, enthusiasm, fear, and so on – because these affective states are short-lived, action-oriented, connected to distinct facial expressions, and typically dependent on real-time environmental responses. However, in that case, they are typically public expressions, and so by their very nature privacy is relinquished with respect to them.

What about states in which my thoughts have linguistic content? This appears to be the gold standard for mind reading. Of course nothing of the sort is, at this point, on the technical horizon, and even were it to be, legal impossibilities would bar general deployment of such a powerful machine. Still, it is nevertheless a useful exercise to consider what is ethically at stake in such a possibility. Would marketers be tempted to use a machine that secretly revealed what a shopper was thinking, in all its detail? Probably, but they would do so at their own peril. To show why consider work done by Thomas Nagel who explored what might happen in case, we lost all reticence.

> A and B meet at a cocktail party; A has recently published an unfavorable review of B's latest book, but neither of them alludes to this fact, and they speak, perhaps a bit stiffly, about real estate, their recent travels, or some political development that interests them both. Consider the alternative:
>
> B: You son of a bitch, I bet you didn't even read my book, you're too dimwitted to understand it even if you had read it, and besides you're clearly out to get me, dripping with envy and spite. If you weren't so overweight I'd throw you out the window.
>
> A: You conceited fraud, I handled you with kid gloves in that review; if I'd said what I really thought it would have been unprintable; the book made me want to throw up – and it's by far your best. (Nagel 1998, p. 12).

The analogical significance of this example to neuromarketing is not immediately obvious until one connects the failure of social civility to the potential failure of market communication. Perhaps Nagel overplays things a bit, but the general philosophical point is that privacy provides protective armor for the conventions of civil life. Removing that armor leads not just to disruptive losses in social communication such as this one, but more serious losses to a range of liberal conventions that permit social life to proceed and flourish.

Applying a version of Nagel's take on this in the face-to-face market context presents a dilemma. Suppose the salesperson knows what the consumer is thinking. Then either the consumer knows this or not. In the case where the consumer knows that the salesperson knows what he is thinking, conversational interactions would be unstable, strange, and unpredictable in content. As a communication strategy, this short circuiting of the norms governing selling behavior seems counterproductive to say the least. On the other horn of the dilemma, the consumer does not know that the salesperson is able to determine exactly what he is thinking in real time. On this horn, a flagrant breach of privacy occurs, a kind of voyeurism motivated by a commercial imperative. In other words, the admittedly (currently, and perhaps even in principle) technically impossible gold standard for mind reading at the store

is either self-defeating or grossly and unarguably unethical, and no properly morally informed and motivated jurisdiction is going to provide statutory permissions to such an institution.

Privacy

Consider now (1) above, namely, what is really at stake in questions of privacy. The definition of privacy is contestable and not just because it is contested, but because two of the salient conditions of privacy – separateness from others (of person, and of personal information) and control over the public-personal boundary – are apparently at odds. The problem is that the use of "privacy," and cognate expressions, carries with it the meaning of both conditions, but logically, one condition may be satisfied and not the other. For instance, a person might voluntarily give up all privacy. In this situation there is no separation from others (and seemingly no privacy), yet full control (and so seemingly privacy is exercised). On the other hand, a neglected elderly person might remain free from intrusion (a condition of privacy), when being left in such an isolated condition is indicative of his loss of control over his interaction with others, his control over the private-public boundary. Resolving this tension is no easy matter, but perhaps it should be said, for present purposes, that privacy is personal isolation, of a certain kind: the kind that occurs when an individual chooses her isolation relatively unconstrained. If this is right, in a privacy violation one objects to another person (or institution) actively perceiving his person (or information), and by that means subjugating him to a motive of theirs that he cannot share (Matthews 2008).

The seriousness of a breach of privacy usually depends on the potential for harm. In the US state of Ohio, Beverly Dennis, a 57-year-old who lived alone, received a letter from a stranger who detailed extensive personal knowledge about her. The letter writer then outlined a range of explicit sexual fantasies he had about Dennis in relation to her preferred brand of hand lotion. It turned out the letter writer was Hal Parfait, a Texas prisoner who had been convicted for invading another woman's home, threatening her children and raping her. Parfait had accessed the personal details of Dennis because Metromail, a data collection agency, had subcontracted the task of inputting personal data – Dennis had earlier completed a discounts coupon – to the Texas Department of Corrections.

The important question for brain privacy, then, is this: what (special) harm may come when neuromarketers obtain my brain information in the non-consenting case? In answering this question, let's assume they have a mind reading device that cannot interpret my linguistic stream of consciousness and cannot determine my background mood. But it can, let's say, tell, in real time, when I am indecisive because of price; it can pick with a 20 % error certain specific emotions; it can sense my loyalty to certain brands (or types of brands), and so on. It can, then, apparently, provide limited help to sales staff in framing their pitch to me.

Our analysis of privacy helps here. As a consumer I have tacitly consented to a situation in which the seller is trying to sell me something I want or to convince

me that I want something (even if perhaps I don't *really* want it). In that case it is just false that the neuromarketer is subjugating me to a motive of theirs I cannot share. My motive in that context is up for grabs. Ceteris paribus, I am not likely to be harmed by the fact that the scanner has delivered information to the salesperson that may make the difference between my buying and not buying. It is not as though the scanner can read my mind and find out my address. (The Dennis case above is in a different league.) In terms of viewing my person, what we object to in perceptual privacy (think here of cases of voyeurism) is that another person has impermissibly viewed my body. The privacy-relevant object here is my self-presentation, which in this context is essentially a presentation of who I am, at the surface. In the case where someone became voyeuristically motivated to view my brain scans, my resentful response that this violated my privacy seems entirely misplaced.

A final point about privacy does express a residual worry. It is one thing to scan the brain of a customer on one occasion, quite a different matter to repeatedly scan the same person's brain over a period of time, thereby building up a *bank* of data. Helen Nissenbaum (1998) provides the locus classicus for the worry being expressed here. Observances of digitally enabled capacities for incrementally obtained information lead to losses of privacy over time, or what she calls the problem of privacy in public. She writes:

> I have in mind here, the principle of contextual integrity and the principle that no information is genuinely "up for grabs", available for purposes such as aggregation, profiling, and data mining. These principles offer criteria for discriminating from among the various forms of public surveillance and record-keeping those that constitute moral violations of privacy and those that do not. (p. 596).

The potential threats to privacy represented by neuromarketers who hold brain information are in this sense no different from any other kind of threat to private information. As such, one should not downplay the threat and nor should one imagine that issues of brain privacy are sui generis. The right response, then, is to assimilate these concerns to the ordinary ethical concerns we have in other sensitive areas. For instance, neuromarketers should be subject to privacy protocols just as medical practitioners are.

Conclusion

The arguments here lead to a mildly skeptical approach to the claims of neuromarketers regarding their capacity for generating valuable data for commercial decision-making. What has been claimed is that to the extent that neuromarketers do succeed in their endeavors, a commensurate ethical concern is raised mainly in relation to questions of privacy, and that is because what is found most offensive in this context is the aggregation of personal data and a loss of control over what may be done with it. These legitimate worries give rise to the need to monitor those neuromarketing practices which risk this loss of control. But theoretically these are not special worries raised by neuromarketing.

Cross-References

- ▶ Brain Research on Morality and Cognition
- ▶ Ethics of Neuromarketing: Introduction
- ▶ Historical and Ethical Perspectives of Modern Neuroimaging
- ▶ Human Brain Research and Ethics
- ▶ Mind Reading, Lie Detection, and Privacy
- ▶ Neuroimaging Neuroethics: Introduction

References

Ambler, T., & Burne, T. (1999). The impact of affect on memory of advertising. *Journal of Advertising Research, 39*(2), 25–34.

Ambler, T., Ionnides, A., & Rose, S. (2000). Brands on the brain: Neuro-images of advertising. *Business Strategy Review, 11*(3), 17–30.

Baumeister, R. F., Bratslavsky, E., Muraven, M. A., & Tice, D. M. (1998). Ego-depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*, 1252–1265.

Bhatia, T. K. (2000). *Advertising in Rural India: Language, marketing communication, and consumerism*. Tokyo: Tokyo Press.

Canli, T. (2006). When genes and brains unite: Ethical implications of genomic neuroimaging. In J. Illes (Ed.), *Neuroethics: Defining the issues in theory, practice, and policy* (pp. 169–183). Oxford: Oxford University Press.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods, 118*, 115–128.

Ehrlinger, J., Johnson, K., Banner, M., Bunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behaviour and Human Decision Processes, 105*(1), 98–121.

Erk, S., Spitzer, M., Wunderlich, A., Galley, L., & Walter, H. (2002). Cultural objects modulate reward circuitry. *Neuroreport, 13*(18), 2499–2503.

Farah, M. J., Smith, E. M., Gawuga, C., Lindsell, D., & Foster, D. (2008). Brain imaging and brain privacy: A realistic concern? *Journal of Cognitive Neuroscience, 21*(1), 119–127.

Fischbach, R., & Mindes, J. (2011). Why neuroethicists are needed. In Judy, J. & J. S. Barbara (Eds.), *Oxford handbook of neuroethics* (pp. 343–376). Oxford: Oxford University Press.

Gazzaniga, M. S. (2006). *The ethical brain*. New York: The Dana Press.

Hubert, M., & Kenning, P. (2008). A current overview of consumer neuroscience. *Journal of Consumer Behaviour, 7*, 272–292.

Kennett, J., & Matthews, S. (2008). What's the buzz? Undercover marketing and the corruption of friendship. *The Journal of Applied Philosophy, 25*(1), 2–18.

Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007). Neural predictors of purchase. *Neuron, 53*(1), 147–156.

Kotler, P., & Keller, K. (2012). *Marketing management*. Boston: Prentice Hall.

Lee, N., Broderick, A. J., & Chamberlain, L. (2007). What is neuromarketing: A discussion and agenda for future research. *International Journal of Pyshcophysiology, 63*, 199–204.

Levy, N. (2007). *Neuroethics*. Cambridge: Cambridge University Press.

Levy, N. (2009). Neuromarketing: Ethical and political challenges. *Etica & Politica/Ethics and Politics, XI*(2), 10–17.

Lewis, D., & Brigder, D. (2005). Market researchers make increasing use of brain imaging. *Advances in Clinical Neuroscience and Rehabilitation, 5*(3), 36–37.

Matthews, S. (2008). Privacy, separation, and control. *The Monist, 91*(1), 130–150.

McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., & Montague, P. R. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron, 44*, 379–387.

Miller, S., & Selgelid, M. J. (2009). *Ethical and philosophical consideration of the dual-use dilemma in the biological sciences*. Dordrecht: Springer.

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114–117.

Murawski, C., Harris, P. G., Bode, S., Domínguez, D. J. F., & Egan, G. F. (2012). Led into temptation? Rewarding brand logos bias the neural encoding of incidental economic decisions. *PLoS ONE, 7*(3), e34155. doi:10.1371/journal.pone.0034155.

Murphy, E. R., & Greely, H. T. (2011). What will be the limits of neuroscience-based mindreading. In Judy, J. & J. S. Barbara (Eds.), *Oxford handbook of neuroethics*. Oxford: Oxford University Press.

Murphy, E., Illes, J., & Reiner, P. B. (2008). Neuroethics of neuromarketing. *Journal of Consumer Behaviour, 7*, 293–302.

Nagel, T. (1998). Concealment and exposure. *Philosophy & Public Affairs, 27*(1), 3–30.

Nissenbaum, H. (1998). Protecting privacy in an information age: The problem of privacy in public. *Law and Philosophy, 17*, 559–596.

North, A. C., Hargreaves, D. J., & McKendrick, J. (1999). The influence of in-store music on wine selections. *Journal of Applied Psychology, 84*, 271–276.

Nutt, D. (2012). *Drugs: Without the hot air*. Cambridge: UIT.

Nwahukwu, S. L. S., Vitell, S. J., Gilbert, F. W., & Barnes, J. H. (1997). Ethics and social responsibility in marketing: An examination of the ethical evaluation of advertising strategies. *Journal of Business Research, 39*, 107–118.

Packard, V. (1957). *The hidden persuaders*. Brooklyn: IG.

Peters, J. F. (Ed.). (2005). *Are your thoughts your own?: "Neuroprivacy" and the legal implications of brain imaging*. The Committee on Science and Law. New York. http://www.nycbar.org/pdf/report/Neuroprivacy-revisions.pdf

Plassman, H., O'Doherty, J., & Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of National Academy of Sciences of the United States of America, 105*(3), 1050–1054.

Plassman, H., Ramsoy, T. Z., & Milosavljevic, M. (2012). Branding the brain: A critical review and outlook. *Journal of Consumer Psychology, 22*(1), 18–36.

Thompson, C. (2003, October 25). There's a sucker born in every medial prefrontal cortex. *New York Times Magazine*. http://www.nytimes.com/2003/10/26/magazine/26Brains.html?src=pm&pagewanted=1

Vohs, K. D., & Faber, R. (2003). Self-regulation and impulsive spending patterns. In P. A. Keller & D. W. Rook (Eds.), *Advances in Consumer Research*, *30*(1), 125–126.

Weisberg, D. S., Keil, F., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477.

Wilson, M. R., Gaines, J., & Hill, R. P. (2008). Neuromarketing and consumer free will. *The Journal of Consumer Affairs, 42*, 389–410.

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality & Social Psychology Bulletin, 31*(1), 121–135.

Zak, P. J., Kurzban, R., & Matzner, W. T. (2004). The neurobiology of trust. *Annals of the New York Academy of Sciences, 1032*, 224–227.

Zak, P. J., Kurzban, R., & Matzner, W. T. (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behavior, 48*, 522–527.

Ethics of Implicit Persuasion in Pharmaceutical Advertising

104

Paul Biegler, Jeanette Kennett, Justin Oakley, and Patrick Vargas

Contents

Abstract

Direct to Consumer Advertising of Prescription Pharmaceuticals (DTCA) is a controversial practice permitted only in the United States and New Zealand. Central to why all other nations ban DTCA is concern about its capacity to impart complete, balanced, and accurate information that guides effective consumer decisions. Yet the debate has, thus far, paid scant attention to how implicit or unconscious persuasion in DTCA might influence consumer attitudes toward advertised drugs. In this chapter, one means of implicit persuasion, evaluative conditioning, is argued to have deleterious effects on the autonomous agency

P. Biegler (✉) • J. Oakley
Centre for Human Bioethics, School of Philosophical, Historical and International Studies Monash University Faculty of Arts, Clayton, VIC, Australia
e-mail: paul.biegler@monash.edu; justin.oakley@monash.edu

J. Kennett
Department of Philosophy, Macquarie University, Sydney, NSW, Australia
e-mail: jeanette.kennett@mq.edu.au

P. Vargas
Department of Advertising, University of Illinois at Urbana-Champaign, IL, USA
e-mail: pvargas@illinois.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_1,

that DTCA viewers bring to medicine choices and on the wider doctor-patient relationship. These effects suggest implicit persuasion should be given much greater consideration in the development of public policy on the marketing of pharmaceuticals.

Introduction

Direct to Consumer Advertising of Prescription Pharmaceuticals (DTCA) is legal in only two nations, the United States and New Zealand. Two primary concerns underpin the decision of all other countries to outlaw the practice. First, prescription drugs have significant potential to cause harm. Prescription status – a requirement that a doctor authorize and oversee use of a drug – is predicated on, among others, a need to safeguard against drug side effects and toxicity (Brass 2001). Prescription status implies, therefore, that the drug carries heightened risk compared to, for example, medications that can be sold over the counter at a pharmacy.

The second concern is that any (legal) decision to take a prescription medicine occurs within the confines of the doctor-patient relationship. That relationship is governed by well-established ethical and legal principles. Prominent among them is a duty to respect and promote patient autonomy, that is, self-governance. That duty demands patients give free and informed consent to medical treatment, including prescription medicines. Respect for autonomy stresses the integrity of the information that patients rely on to make medical decisions. It is known that exposure to DTCA influences patients' requests for drugs and, as a result, doctors' prescribing behavior (Mintzes et al. 2003). Hence, the quality of product information provided by DTCA can wield formidable influence on the doctor-patient relationship.

Opponents argue that DTCA harms by promoting excessive and sometimes inappropriate pharmaceutical use (Hasman and Holm 2006). In support, there is substantial evidence for the persuasive force of DTCA. On one estimate DTCA returns around $US4 for every dollar spent by pharmaceutical manufacturers (Mintzes 2009). Further evidence suggests that increased sales of advertised medicines are driven by patient requests. People who view DTCA are more than twice as likely to make brand-specific drug requests compared to those in jurisdictions without DTCA (Mintzes et al. 2003). Doctors honor those requests in over 50 % of cases (Kravitz et al. 2005). Yet there is doubt whether improved public health results. Kravitz et al. (2005), for example, showed that patient requests for a brand-specific drug were met in 55 % of instances where the drug was not a recommended treatment, specifically, antidepressants for adjustment disorder. And in a meta-analysis, Gilbody et al. (2005) concluded that while DTCA increased drug prescriptions, no study showed evidence of overall public health benefit.

Opponents further argue that the persuasive intent of DTCA leaves informational integrity as a secondary concern. Advertisers are, on this view, motivated to provide inadequate, incomplete, imbalanced, or misleading drug information if, in so doing, the drug is presented in a more favorable light (Hasman and Holm 2006). This concern is heightened in the United States by the absence of any requirement

for regulators to vet DTCA pre-release. Rather, ads are subject to post hoc scrutiny only if a complaint is made (Food and Drug Administration 2012). The result, opponents argue, is an impoverished understanding among DTCA viewers that militates against autonomous choices and informed consent.

DTCA proponents typically counter that advertising is an important means of promoting awareness of both diseases and their treatment (Hasman and Holm 2006). This is held to be of special significance for the socioeconomically disadvantaged for whom advertising, especially of the broadcast variety, is a primary source of medical information. It is also argued that DTCA acts as a reminder function that enhances compliance of existing patients with their medication regime (Mintzes 2006). Proponents further argue that DTCA bans are an unjustified constraint upon free speech. These grounds formed the basis for legal action challenging the constitutional validity of Canada's prohibition of DTCA (Priest 2007). That process was terminated, however, before any judicial decision was reached.

Yet, to date, the debate has largely ignored consumer psychology research that increasingly shows how advertising persuades outside of awareness. These subtle forms of persuasion may impact the attitudes viewers come to hold about advertised drugs, while escaping current regulation. Further discussion of unconscious or "implicit" persuasion requires some clarification of terms. There are three prominent accounts of implicit persuasion (Chartrand 2005). First, the individual may be unaware of the persuasive stimulus. In subliminal advertising, for example, frames in a commercial are presented too briefly to be consciously processed. Second, the individual may be unaware of the process by which persuasion occurs. As will be detailed, this form of implicit persuasion can occur even when all stimuli are supraliminal, that is, above the threshold of consciousness. Finally, individuals may be unaware of the attitudes that they come to hold as a consequence of the persuasive process. This third possibility is predicated on a dual theory of attitudes (Wilson et al. 2000). On this view, people hold conscious or explicit attitudes that they can report, but also implicit or unconscious attitudes of which they are unaware. People may report, for example, explicit attitudes of racial impartiality yet be gauged, on measures such as the Implicit Association Test, to hold implicit attitudes consistent with racial prejudice (Greenwald et al. 1998).

The following discussion focuses on the second category; implicit persuasion where the individual is unaware of the persuasive process. Subliminal persuasion is not addressed because of uncertainty about its real world effectiveness (Vargas 2008) and because it is almost universally prohibited. Nor are implicit attitudes considered, primarily because of controversy over what implicit measures actually assess, and the degree to which they predict subsequent behavior (van Ravenzwaaij et al. 2011). The focus is, therefore, on persuasion using supraliminal stimuli where the resulting attitude can be reported by the individual, yet where the mode of persuasion remains hidden. The importance of this category is that it covers techniques that are permissible within current regulations and detectable on relatively uncontroversial measures.

A number of techniques fit this classification. For example, the mere exposure effect (Zajonc 1968) describes how repeated viewing of a neutral stimulus can

cause people to like the stimulus. Consistent with mere exposure, repeated viewing of advertised products increases both familiarity with and liking of those products (Fang et al. 2007). Priming describes how advertisements act as hedonically charged cues that enhance the product's mental accessibility and activate striving for them (Chartrand et al. 2008). Framing effects occur when identical information is presented in different ways, causing divergent attitudes to the target object. For example, the description "95 % fat-free" generates greater liking than does "5 % fat" (Kahneman et al. 2000).

Perhaps one of the best researched techniques in this category is evaluative conditioning, a variant of classical Pavlovian conditioning. Evaluative conditioning results from the pairing of a neutral stimulus, for which no special feelings are held, with a valenced stimulus, one that elicits either positive or negative feelings (Jones et al. 2010). With repeated pairing, the feelings and consonant attitude held toward the valenced stimulus pass to the neutral stimulus. In the case of a positive valence stimulus, the result is a positive attitude toward the previously neutral stimulus. In advertising, for example, images of happy smiling people may be used to condition positive attitudes toward a novel product, for example, a new brand of shampoo.

In the language of Pavlovian conditioning, the valenced image is termed the *unconditioned stimulus*, because it elicits positive feelings with no need for further manipulation. The neutral object, the shampoo, is the *conditioned stimulus* that takes on the positive valence of the unconditioned stimulus. The resulting positive attitude is the *conditioned response*.

In what follows, evaluative conditioning is used as a paradigm case of implicit persuasion in DTCA to examine the consequences for the autonomy with which consumers make choices about medicines, their capacity for agency in relation to those choices, and the subsequent impact on the doctor-patient relationship.

Empirical Evidence for Implicit Persuasion via Evaluative Conditioning in DTCA

In this section, evidence is summarized for the persuasive potency of evaluative conditioning, its operation in DTCA, and its implicit nature.

Evaluative conditioning has been demonstrated toward a range of conditioned stimuli including words (De Houwer et al. 1994), odors (Rozin et al. 1998), and human faces (Walther 2002), using various valenced unconditioned stimuli, among them, images (Pleyers et al. 2007), music (Eifert et al. 1988), and candy (Brunstrom and Higgs 2002). While concerns have been voiced that conditioning effects are difficult to replicate (Rozin et al. 1998), a recent meta-analysis of 214 evaluative conditioning studies concluded it to be a robust phenomenon (Hofmann et al. 2010).

In a typical study, Sweldens et al. (2010) used positive valence images to condition positive attitudes toward novel brands of Belgian beers. Smaller images of branded beer bottles were superimposed over a variety of larger positive images, including people water-skiing, sailing, and cuddling. Positive images were sourced from the International Affective Picture System (IAPS), a database where the

valence of each image has been systematically calibrated (Lang et al. 2008). Participants exposed to positive images reported more positive attitudes toward the beers than did those exposed to neutral valence images, for example, people reading a newspaper or napping on a subway.

While evaluative conditioning is strongly supported by research in social psychology, proof of its operation in DTCA faces some empirical hurdles. First, a "process pure" evaluative conditioning experiment involves repeatedly pairing a conditioned stimulus with unconditioned stimuli whose valence is known. Varying the valence of the unconditioned stimulus and noting consonant alterations in the induced attitude permit ascriptions of persuasive potency to the unconditioned stimulus. Yet, as De Houwer (2009) has noted, real world ads pair the product, the conditioned stimulus, with a variety of valenced stimuli including imagery and music, and also with propositional informational content. The multiplicity of attitude influences is an impediment to discerning the relative contribution of conditioning against other persuasive content.

Yet, as Biegler and Vargas (2013) have argued, there exist both conceptual and empirical rationales to accept the existence of evaluative conditioning in DTCA. First, imagery in DTCA relates closely to the positive valence pictures featured in the IAPS. For example, an ad for Lipitor (Pfizer 2010) features a man petting a dog and leaping from a jetty before swimming in a sparkling mountain lake. It is not coincidental that cute animals and majestic natural scenery figure prominently in the most positive valence images of the IAPS. There is also a database of sounds whose valences have been calibrated, the International Affective Digitized Sounds (IADS) (Bradley and Lang 2007). The most positive valence sounds include the music of Bach and Beethoven, people laughing, and a baby cooing. Uplifting music and the sounds of people enjoying themselves abound in pharmaceutical commercials. For example, an ad for the asthma inhaler Advair (GlaxoSmithKline) has children laughing after blowing bubbles and hitting a piñata at a sun-drenched outdoor party. Guitar music in a major key provides a soaring backdrop.

There is strong prima facie evidence for the existence of positive valence unconditioned stimuli in DTCA. Further, many commercials repeatedly present images of branded drug boxes and logos, which plausibly comprise the conditioned stimulus. The conclusion that evaluative conditioning will occur is compelling. Indeed, in an extensive review, Schachtman and colleagues concur that:

> [In] the case of paired events during advertisements, if the individual changes his or her behavior (attitude change, interest in purchasing the item) in the presence of the product or brand (the conditioned stimulus) as a function of pairings of this conditioned stimulus with an affective stimulus (the unconditioned stimulus), then this behavior can be said to be a conditioned response (Schachtman et al. 2011, pp. 481–482).

An experiment conducted by Smith et al. (1998) provides a supporting empirical rationale. On their construal, a commercial's positive valence elements, including imagery and sound, can combine to elicit positive affect. On this view, multiple stimuli of similar valence can comprise the unconditioned stimulus. Consistent with

this view, Smith et al. (1998) found commercials that elicited broadly positive affect also induced the most positive attitudes toward the featured product.

A remaining question is whether evaluative conditioning is applicable to pharmaceuticals. It might be argued, for example, that pharmaceuticals are not "neutral" stimuli and, therefore, may not be ideal conditioned stimuli. In fact, there is evidence that many connote drug treatments negatively (Benkert et al. 1997). If so, pharmaceuticals may evince differential susceptibility to conditioning compared to say, shampoo, or dishwashing liquid. Addressing this concern, Biegler and Vargas (in preparation) demonstrated positive evaluative conditioning toward a hypothetical influenza drug using valenced imagery from the IAPS. This study suggests that traditional conditioning techniques are indeed applicable to pharmaceuticals, albeit in relation to one category.

To sum, there is good evidence that evaluative conditioning is a robust effect that is operative in DTCA. But is its operation implicit, that is, outside awareness? To support this claim, it is necessary to look more closely at the conduct of evaluative conditioning experiments. A number of the studies cited earlier that successfully conditioned positive attitudes also included demand awareness checks (Brunstrom and Higgs 2002; Rozin et al. 1998; Walther 2002; Sweldens et al. 2010). These measures test to see if participants became aware of the investigators' hypothesis during the experiment. The danger is that awareness may cause participants to behave in line with what they assume to be the "demands" of the experimenter. No study, however, reported significant awareness of the intent to condition attitudes. Admittedly, these experimental cohorts differed from the real world public who view DTCA. Many participants were business or psychology undergraduates. However, these individuals can be expected to have a relatively sophisticated understanding of experimental psychology. Given their broad ignorance of the conditioning process, it is plausible that members of the general public would evince equivalent or lesser awareness of evaluative conditioning in DTCA.

Consumer Autonomy and Justified Beliefs About Material Facts

To understand the potential impact of implicit persuasion, exemplified by evaluative conditioning, on consumer autonomy, it is first necessary to mount a plausible theory of autonomy. The contemporary literature expounds many competing theories. Prominent among them are Hierarchical (Frankfurt 1971; Dworkin 1988), Life Plan (Young 1986), Historical (Christman 1991), Reasons-Responsive (Fischer and Ravizza 1998), and Relational (Mackenzie and Stoljar 2000) accounts. These theories are merely alluded to because the intention is to emphasize a commonality in each that has relevance for the current argument. While each account advances varying specifications for the kinds of desires that best underpin ascriptions of autonomy, each is dependent upon the agent holding a sound epistemology. Thus, it is largely uncontroversial that, should an agent wish to pursue decisions or actions that accord with deeply held desires, goals, or values, a grasp of pertinent facts is a necessary tool.

Perhaps the realm of medical ethics and informed consent has focused most attention on the epistemic requirements for autonomous choice. It is widely held that for patients to make fully informed and autonomous medical decisions, they must understand facts material to the treatment in question. Faden et al. (1986) provide insight into the nature of material facts. They argue that facts are material if they "would be viewed by the actor as worthy of consideration in the processes of deliberation about whether to perform a proposed action" (Faden et al., p. 303). Material facts are, therefore, those the individual considers important in making their decision about treatment. This "particular patient" standard of materiality is supported by legal precedent. In the Australian case Rogers v. Whitaker (1992), a woman sought a largely cosmetic surgical procedure to an eye that had been blinded in a childhood accident. The surgeon failed to warn her of the risk that surgery could rob her of sight in the functioning eye. This idiosyncratic complication, known as sympathetic ophthalmia, has a probability of 1 in 14,000. The complication occurred, and the judge commented that, objectively low odds aside, the patient's view that the risk was material was definitive.

Yet Faden and colleagues also make clear that for autonomy to be deduced, patients must not simply understand material facts, but also believe them (Faden et al., p. 311). Understanding that the physician asserts one's toe is gangrenous, for example, does not satisfy an autonomy standard if one remains committed to the belief the toe is just dirty. Here, understanding does not guarantee the fact of a gangrenous toe will be given due weight in the ensuing deliberations. But if belief is to be the relevant benchmark for autonomy, then epistemology offers two further potential standards: justified or true belief. True belief holds intuitive appeal – it seems obvious that a prospective patient must hold an accurate view of, say, the nature and site of an operation for autonomous consent to be proffered. Yet the true belief standard has drawbacks. A person with dementia may hold beliefs that alternate between the true and the fictional. If consent is obtained when the true belief is held, ought this to be seen as a moment of lucidity? Or is neurological deterioration merely throwing up vague semblances of reality? This is an example of "epistemic luck" which many philosophers eschew as grounds for legitimate knowledge.

In contrast, a justified belief standard emphasizes the reliability of the belief-forming process, where such a process enhances the likelihood that true beliefs will result. Goldman and Olsson (2008) provide an illustrative analogy. They describe a pair of scenarios in which a driver must negotiate two forks in the road to reach a town. In the first scenario, the driver has a reliable satellite navigation unit that, as expected, gives the correct direction at the first fork. In the second scenario, the driver has an unreliable navigation unit that, by chance, gives the correct direction at the first fork. Goldman and Olsson argue that, while drivers in both scenarios may hold a true belief about the correct direction at the first fork, the driver's belief in the first scenario carries greater epistemic value. That superior value stems from the fact that a reliable navigation unit heightens the

probability of a correct decision at the second fork, and ultimately reaching the desired destination. As the authors put it:

> [T]he conditional probability of getting the correct information at the second crossroads is greater conditional on the navigation system being reliable than conditional on the navigation system being unreliable (Goldman and Olsson 2008, p. 28).

This "conditional probability" thesis has relevance for autonomy in that most medical choices require not just one, but a suite of decisions in relation to a range of contingencies. For example, the decision to use a particular medication will require weighing its therapeutic efficacy against its side effects. The decision may also demand a comparison between the medication and alternative medicines, non-pharmacological therapies, or indeed no treatment at all. Should each belief derive from a reliable process, for example, a reading of evidence-based plain language information leaflets, there is a heightened probability of a true belief issuing at each juncture. By contrast, should the belief-forming mechanism be unreliable, a true belief may issue by chance in relation to one contingency, but the prospects are poor that outcome will be repeated. The result is a corruption of the deliberative process.

There are strong reasons to hold that evaluative conditioning in DTCA and other implicit persuasive techniques comprise unreliable belief-forming mechanisms that foster unjustified beliefs about pharmaceuticals. Moreover, many such beliefs will pertain to facts that are plausibly material to viewers with the relevant illness, such as the drug's safety and efficacy. Given a justified belief standard for autonomy, evaluative conditioning and related techniques pose a threat to autonomous drug choices. This threat is elaborated now.

Implicit Persuasion and Consumer Autonomy

Evaluative conditioning in DTCA is an inherently unreliable progenitor of beliefs about drug facts. To see this, it is first necessary to outline how evaluative conditioning influences belief. Evaluative conditioning employs valenced unconditioned stimuli to modify attitudes to a neutral conditioned stimulus. On a "tripartite theory" of attitude structure, the resulting attitude comprises affective, cognitive, and behavioral elements (Bizer et al. 2003). On this view, the transfer of positive valence from unconditioned to conditioned stimulus is accompanied by consonant beliefs about, and behavioral intentions toward, that stimulus. For example, if images of people enjoying themselves were used to condition positive attitudes toward an automobile, the tripartite theory predicts those attitudes to encompass positive beliefs, for example, that its components are of higher quality, and positive intentions, for example, a favorable disposition to purchase it.

There is empirical evidence of the impact of evaluative conditioning on belief. For example, Krosnick et al. (1992) used valenced imagery to condition positive and negative attitudes toward images of a woman studying and shopping. Participants conditioned to hold positive attitudes believed the woman to be friendlier, kinder, fairer, and more honest. Extending this research, Biegler and Vargas

(in preparation) found that participants conditioned (with positive imagery) to hold positive attitudes toward a hypothetical pharmaceutical believed it to be significantly safer, more effective, and more beneficial than did participants conditioned (with negative imagery) to hold negative attitudes toward the pharmaceutical.

Yet, it must be remembered that while the valence of an image used as an unconditioned stimulus influences belief about the conditioned stimulus, that valence bears no substantive relationship with characteristics of the conditioned stimulus. For example, the most positive valence image in the IAPS features three puppies perched expectantly on a wall. Should favorable beliefs emerge toward a co-presented pharmaceutical as a result of pairing with that image, there is no basis to conclude that such an effect will promote true beliefs. This is especially so should valenced stimuli be used with the intention of promoting favorable beliefs, at the expense of their veracity.

It may be countered that DTCA contains negative valence content and that positive conditioning plausibly corrects for this. For example, there is a requirement for extensive disclosure of side effects ranging from, for example, nausea and headache to life-threatening effects such as throat swelling or liver failure. Further, there is evidence that negative terms can act as unconditioned stimuli that can produce negative attitudes (De Houwer et al. 1994). On this view, positive imagery may be a necessary epistemic antidote.

It is conceded that such an effect may occur. The issue, however, is whether evaluative conditioning promotes true beliefs across a range of decisional instances. A person considering use of a medication, for example, an antidepressant, will be exposed to multiple advertisements for a range of drugs. Should evaluative conditioning occur in every ad, there is little chance that it will engender accurate beliefs in each case. It must be remembered that the primary intention of the technique is to persuade, not instill true belief, and so the latter will occur only haphazardly.

The relevance for autonomous choice is that such an unreliable influence upon beliefs about salient properties such as drug effectiveness and safety will surely lead to unjustified beliefs. And given that such properties are material to most who consider using a medication, autonomous choice is set back. Specifically, because unreliable belief-forming mechanisms heighten the chance of holding false beliefs, there is a lessened probability that drug choices will be in accord with the agent's values. Given the intense efforts in recent years to ensure that patients' decisions reflect a robust standard of autonomy, this effect of DTCA must give cause for serious circumspection about its permissibility, at least in its current form. That conclusion, however, faces an important objection that stems from the impact of recent psychological research on conceptions of human agency.

Implicit Persuasion and Agency

The argument thus far assumes a picture of human agency that sees agents as capable of rational deliberation and choice under conditions of sufficient information and absence of coercion or hidden persuasive factors. But what if this is hardly ever the case? And what if this picture of agency is mistaken?

One contemporary response to the concerns raised above about the effects on patient autonomy of implicit persuasion in DTCA is a skepticism about rational agency itself. This skepticism arises from a vast array of psychological data – some already referred to – demonstrating that we are largely unaware of many rapid and automatic influences on individual choice, judgment, and belief. These data suggest that, notwithstanding our beliefs to the contrary, individuals really have little idea of why they make the choices they do. Bargh and Chartrand (1999), for example, say:

> Most of a person's everyday life is determined not by their conscious intentions or deliberate choices but by mental processes that are put in motion by features of the environment and that operate outside of conscious awareness and guidance (Bargh and Chartrand 1999, p. 462).

For example, when subjects in a shopping mall were asked to make a choice between identical items on a stand, they showed a marked preference for the item on the right. But when asked to provide reasons for their choice, no subject mentioned this. Instead they produced explanations that focused on the supposed superior features of the product they chose (Nisbett and Wilson 1977, reported in Carruthers 2005, pp. 142–143).

According to a dual process model of cognition, many judgments that people make are the product of automated processing that occurs beneath the level of awareness. Indeed skeptics about the traditional picture of rational autonomous agency claim that *reasoning is not for what we think it is for*. It is not a vehicle for arriving at truth and securing autonomy in decision-making. Jonathan Haidt (2001, 2012) argues that its main function is to generate post hoc justifications for the intuitive judgments a person makes and to secure social agreement.

A skeptical view of the scope and role of conscious reasoning seems to make it more urgent to regulate persuasion in pharmaceutical advertising that operates below awareness, produces favorable product attitudes, and is not based on information. But there are at least two reasons to doubt this would make measurable difference to the quality of people's decisions. First, as Bargh and Chartrand (1999) note, the environmental influences on choice are numerous and varied. Our judgments have been found to be influenced by the weather (Schwarz and Clore 1983), by mood (Yuen and Lee 2003), by whether we are sitting at a clean or dirty desk (Haidt and Bjorklund 2008), and even by reading a report which uses plural rather than singular pronouns (Gardner et al. 1999). Positive imagery in advertising may have some influence on the beliefs people hold and the decisions they make. But it does not obviously follow that it makes their decisions worse or less autonomous than they would have been in the absence of such imagery. This is because the decision-making process is subject to many other unconscious influences which are likewise irrelevant to the properties of the product being evaluated.

Second, processing of purely *propositional* information is also subject to a range of cognitive and motivational biases which lead to incorrect beliefs. Persons are capable of more effortful and explicit controlled processing which can correct some of the biases to which we are subject. Kahneman (2011), however, argues that the controlled processing system is "lazy" and prefers to accept the deliverances of the

automatic system rather than to interrogate them. Jonathan Haidt (2012, pp. 83–88) surveys a range of evidence to support the view that when we do reason, the search for evidence is biased toward the conclusion we want to reach. Motivated reasoning, as it is known, selectively directs our attention to a subset of relevant information (Kunda 1990). Haidt argues that often a single piece of supporting evidence is sufficient to give us permission to believe what we want. For the purposes of considering the ways in which patients might process information in pharmaceutical advertising, two studies cited by Kunda (1990) stand out. First, subjects diagnosed as having a (fictitious) enzyme deficiency rated the condition less serious and the test less accurate than did subjects diagnosed as not having it (Ditto et al. 1988). Second, Kunda (1987) conducted a study in which subjects read a fictitious study claiming that caffeine was risky for women. Women with a high caffeine intake found the science less persuasive than women with a low caffeine intake or men. "Only subjects who stood to suffer serious personal implications if the article were true doubted its truth" (Kunda 1990, p. 489).

But even without such motivational influences on information processing, there is evidence that controlled effortful processing may not be the best way to arrive at important decisions.

> Conscious thought has shortcomings that can prevent sound decision making. First of all, conscious thought can lead to suboptimal weighting of the importance of aspects of different choice alternatives. In addition, because consciousness has *low capacity*, conscious thought often leads people to take into account only a limited subset of information at the expense of other information that should be taken into account when making a decision (Dijksterhuis and van Olden 2006, p. 628).

Lisa Bortolotti (2011) cites evidence suggesting that attitudes and choices arrived at via reason-giving are less optimal with respect to expert opinion and more vulnerable to evidence manipulation than those made without giving reasons (e.g., Wilson and Schooler 1991). This kind of evidence has been taken by both Jonathan Haidt (2001, 2012) and John Doris (2009) to show that the picture of persons as rational reflective agents is just false. It follows that *reasoning* (at least individually) is not a particularly reliable way to arrive at conclusions about what we have reason to do.

Thus, persons may not be the kinds of agents that they are assumed to be and careful reflection may not be a reliable way to get at the facts relevant to a decision. Further, decisions may be influenced by multiple and random persuasive factors in the environment of which individuals are not and cannot be aware, no matter how hard they try. From this, it might be argued that evaluative conditioning and other forms of implicit persuasion in DTCA do not leave patients in a worse epistemic and agential position than before. This may be especially the case given that mandated presentation of negative side effects in DTCA could lead to inflated assessments of risk.

Individuals have been found to overestimate the risk and prevalence of certain conditions. In addition to the powerful influence of framing effects on assessments of the risk and benefit of policies and actions – a scenario framed in terms of lives lost is evaluated more negatively than an identical scenario framed in

terms of lives saved (Tversky and Kahneman 1981) – the *alarmist bias* holds that "the worst possible scenarios loom large in people's minds, distorting their risk perceptions and their behaviours" (Kuran and Sunstein 1999, p. 706). As noted earlier, evaluative conditioning might correct such a distortion of risks and leave the patient in a better epistemic position than would the presentation of propositional content alone.

But evidence about the limits of rational agency and the need to counter irrational assessments of risk made by patients exposed to DTCA cannot constitute an argument for permitting implicit persuasion in DTCA. Rather it seems to constitute an argument against permitting DTCA at all.

Maybe in some cases, knowing the true reasons for, or causes of, our choices is not important. The fact that consumers tend to choose products on the right side of a display, and not because they are superior in the ways the consumer might reason, does not really matter. But, in many instances, getting it right does matter and so being aware of cognitive limitations and factors that shape deliberation outside awareness might be rather important. Choosing the medication most appropriate to a medical condition is something agents surely have reason to do and, as already argued, this choice ought to be responsive to facts about the medication and not to irrelevant features of an advertisement. However, given that human agency is systematically vulnerable to epistemically misleading influences, there is a need to consider what policies and practices support good decision-making, that is, decision-making in accordance with material facts.

One approach to the problem is social design. We can introduce nudges of various sorts to move us toward better decisions (Thaler and Sunstein 2009). Such nudges are common in health-related areas. They can include regulatory measures such as price signals, plain packaging of dangerous products, and warning labels. They also include government-sponsored advertising campaigns – for example, highlighting the dangerous effects of speeding, smoking, or excessive drinking. Such sponsored health campaigns, of course, often use similar persuasive techniques to those discussed here. They pair positive imagery with positive choices and negative imagery with dangerous choices. Do they undermine consumer agency or autonomy in so doing? If it can be argued that this use of implicit persuasion is benign because it nudges the consumer into a more accurate appreciation of the material facts relevant to a decision, could the same argument be available to pharmaceutical companies?

The obvious response to such claims is that there is no evidence that DTCA leaves us in a better epistemic and agential position and reasons to think that it does not. Whereas government health campaigns are motivated by paternalistic considerations, the goal of pharmaceutical advertising is neither paternalistic nor concerned with support for epistemic agency. Companies are motivated to increase sales and improve profits for shareholders. DTCA does so in part by implicit persuasive techniques and in part by providing explicit information that the evidence suggests consumers will be ill-equipped to evaluate. Its proponents claim that it enhances our agency by bringing more options to consumer attention. But if these options cannot be properly evaluated, consumers may be worse off.

A second approach to better decision-making is a reliance on experts. The account of the limitations of human agency and decision-making is at odds with an ideal of rational autonomous decision-making that is dominant in philosophy and bioethics. However, there are exceptions to this picture. Expert decision-making conforms more closely to ideals of agency in that it is less subject to the kinds of biases and distortions that plague the nonexpert, and more responsive to reasons and evidence. Lisa Bortolotti (2011) cites the following features of expertise identified by Hutton and Klein (1999). Experts in a domain are better able to perceive patterns; through years of experience, experts acquire the ability to perceive relevant features of the situation; their performance is virtually error-free; they display superior memory in their domain of expertise, have a deeper understanding of the problem to solve (e.g., they catch on to the causal mechanisms); they have a better understanding of their own limitations and an ability to catch themselves when they commit errors. It is important, however, to note that decision-making by experts does not usually require explicit reflection; it may be for the most part automatic or intuitive. Further, experts are by no means immune to the kinds of biases detailed here (Drew et al. 2013).

It should be apparent that many, or even most patients are not expert agents with respect to their medical condition or to pharmacology. If their decision-making in this important area is to count as autonomous, their agency must be supported by someone who *is* an expert and is bound to consider their best interests. Medical practitioners and pharmacists are the most appropriate candidates for this role. Pharmaceutical companies and advertisers are not.

At this point, proponents of DTCA may point to the important fact that a patient may become aware of a product for their condition via advertising, yet be unable to access it without a prescription. So their choice or preference must be discussed with a medical practitioner and their prescription must also be filled by a pharmacist. There will be an expert gatekeeper, and so their agency will receive appropriate scaffolding. It is appropriate then, to consider whether this is the case. How might the persuasive effects of DTCA on consumers impact the doctor-patient relationship?

Implicit Persuasion and the Doctor-Patient Relationship

A number of empirical studies have shown that DTCA significantly increases inappropriate prescribing (e.g., Mintzes et al. 2002; Murray et al. 2003; Gilbody et al. 2005). This increase seems largely because doctors working in DTCA environments face greater demands from patients for inappropriate medication, and because many doctors report that they find it difficult to resist such demands. As will be detailed, there is good reason to think that evaluative conditioning intensifies the impact of DTCA on inappropriate prescribing through such mechanisms. DTCA of pharmaceuticals also arguably undermines ethical medical practice in more subtle ways, as the impact of DTCA threatens to redefine therapeutic doctor-patient relationships as consumer relationships.

What do we know about the impact of pharmaceutical DTCA on doctor-patient relationships? Pharmaceutical companies invest heavily in DTCA, because, as noted above, it significantly increases sales. However, its mechanisms of influence over clinical practice have not been well understood as there was, historically, little systematic empirical research to substantiate various claimed effects of DTCA on doctor-patient relationships. More recently, however, a number of large-scale empirical studies have investigated those questions. For example, Murray et al. (2003) found that 74 % of the 535 US doctors they surveyed had seen patients in the previous 12 months who discussed DTCA drug information with them. Forty-eight percent of those patients were doing so because they wanted a change in medication, which in almost half of cases the doctor regarded as clinically inappropriate. The study also found that doctors "often seem to acquiesce to [clinically inappropriate] requests so long as the patient is not harmed" (Murray et al. 2003, p. 521). The researchers concluded that "DTCA results in patients making almost as many inappropriate requests as appropriate ones (Murray et al. 2003, p. 521)." In their 2004 study of patients, Murray and colleagues found that 161 of the 226 respondents who said they had discussed DTCA at a medical consultation requested a specific intervention or medication after viewing DTCA (Murray et al. 2004). Seventy of these respondents reported not receiving that intervention or medication from their doctor, despite requesting it (Murray et al. 2004, pp. 13–15). These findings raise ethical concerns about DTCA undermining patient autonomy in decisions about medicines, and jeopardizing beneficent action by doctors toward patients.

Similar concerns are raised by Mintzes et al. (2003). They surveyed 78 primary care physicians and 1,431 patients in Sacramento and Vancouver[1] about patient requests for specific drugs after exposure to DTCA for those drugs. The study found that 7.2 % of the Sacramento patients requested DTCA-advertised drugs, compared with 3.3 % of Vancouver patients (Mintzes et al. 2003, pp. 408–409). These findings are notable because patients' self-reported exposure to pharmaceutical DTCA was significantly higher in Sacramento than in Vancouver. Further, once a patient requested a drug, both the US and Canadian physicians were likely to fulfill such requests, but were particularly likely to do so when the request was for one or more DTCA drugs. The researchers reported that "In Sacramento 80 % of patients who requested prescriptions received them, as compared with 63 % in Vancouver. . .., Indeed, for patients requesting DTCA drugs, the odds of receiving a prescription. . .were 16.9 times those of patients who did not request a medicine" (Mintzes et al. 2003, p. 409). The researchers concluded that: "If DTCA opens a conversation between patients and physicians, that conversation is likely to end

[1]Although DTCA of prescription pharmaceuticals is prohibited in Canada, there has nevertheless been a proliferation of "reminder" or "branded" advertisements on television by Canadian pharmaceutical companies. These ads mention a brand name of a drug without specifying what condition the drug treats (see Silversides 2001) and are tolerated by Canadian regulators. Canadians may also view DTCA via US-based cable TV and the Internet.

with a prescription, despite frequent physician ambivalence about treatment choice. And the greater the patient's exposure to advertising, the more likely such a conversation will occur" (Mintzes et al. 2003, p. 412).

The first systematic review of the benefits and harms of DTCA concluded that, overall, "Direct to consumer advertising is associated with increased prescription of advertised products and there is substantial impact on patients' requests for specific drugs and physicians' confidence in prescribing" (Gilbody et al. 2005, p. 246). The same study corroborated the view that physicians often capitulate to patient demands for the advertised drug, despite physicians' misgivings about the drug in question. The strong influence of DTCA on patients making "brand-specific requests" following exposure to DTCA is also supported in the study of antidepressant requests cited earlier (Kravitz et al. 2005). The researchers trained actors to perform standardized patient roles to investigate the effects of patients' DTCA-based requests (and particularly brand-specific requests) for antidepressant drugs on doctors' prescribing behavior. Of 149 of these patients presenting with depression, 27 % of those who requested a specific brand of antidepressant received that drug. Of 149 patients presenting with adjustment disorder, for which medication is not a recommended treatment, 36 % of those patients had their brand-specific request fulfilled. The researchers concluded that "Brand-specific requests had a differentially greater effect in adjustment disorder compared with major depression. This supports the hypothesis that DTC advertising may stimulate prescribing more for questionable than for clear indications" (Kravitz et al. 2005, p. 2000). The authors also comment that "If patients can sway physicians to prescribe drugs they would otherwise not consider, physicians may not be the stalwart intermediary that the law assumes" (Kravitz et al. 2005, p. 2000). So, patients in DTCA environments seem to become more demanding of their doctors regarding prescription pharmaceuticals, and doctors often seem to give in to the pressures they feel from patients to prescribe a particular drug (see also Spurgeon 1999).

The effects of DTCA on inappropriate prescribing seem magnified when DTCA employs techniques such as evaluative conditioning to condition favorable attitudes toward the advertised drug. As argued in the first section, DTCA using evaluative conditioning undermines patients' ability to take an informed view about what medicines are best for them (see also Biegler and Vargas 2013). But there is also evidence that evaluative conditioning in DTCA influences patients to make brand-specific requests from their doctor.

A May 2012 study randomized 140 US undergraduate students to view either an actual TV commercial for a sleeping pill or read and hear a transcript of the ad's voiceover and printed statements (Biegler et al. unpublished study). The ad utilized positive valence imagery consistent with that used in evaluative conditioning studies. It also included a list of side effects, some of which are potentially serious. One dependent measure gauged the likelihood the participant would request the sleeping pill from a doctor if insomnia were experienced. Responses were recorded on a 7-point scale where 1 = extremely unlikely to request and 7 = extremely likely to request the pill from a doctor. Ad viewers' average rating of their likelihood of

requesting the sleeping pill was 4.47, compared with 4.13 (i.e., just above neutral) for transcript viewers. This difference was marginally significant, $F(1, 136) = 3.67$, $p = .057$. However, this difference was entirely driven by participants who were not native English speakers ($n = 51$); the marginal effect of ad condition on participants' likelihood of requesting the sleeping pill from their doctor depended on whether they were native English speakers, $F(1, 136) = 6.16$, $p = .01$. Non-native English-speaking ad viewers' average rating of their likelihood of requesting the sleeping pill from their doctor was 5.24, compared with 4.10 (just above neutral) for non-native English-speaking transcript hearers.

These findings support the persuasive power of the commercial's "nonpropositional content," that is, music, imagery, and voiceover tone, for example, that does not comprise an explicit claim about the drug. Nonpropositional content was differentially more persuasive for those less able to evaluate the ad's explicit, propositional claims. The findings buttress the ethical concerns already raised about the capacity of some DTCA viewers to accurately gauge an advertised medicine's properties and whether their subsequent requests for those drugs (often met by doctors) are in their best interests. Proponents cannot simply assume that DTCA is an unqualified good in terms of raising patient awareness of (and compliance with) medication, regardless of what method of advertising is used.

There is also good reason to hold that DTCA distorts the therapeutic nature of doctor-patient relationships. That is, DTCA threatens to redefine doctor-patient relationships as *consumer* relationships by altering the proper governing conditions of those relationships. Those governing conditions are demonstrated by what doctors *prioritize* in their clinical decisions, and by their *reasons* for prioritizing those considerations. For instance, a doctor who regularly prescribes drugs because they are demanded by patients responding to DTCA, whether or not the drug is optimal, demonstrates a consumer relationship rather than a therapeutic relationship with those patients (see Oakley 2012, 2014). The use of evaluative conditioning in DTCA heightens these concerns. This is because its persuasive potency will likely increase the frequency of such patient requests. Moreover, data on the resistance to extinction of evaluative conditioning suggests its presence in DTCA will make it more difficult for doctors to counter patients' ad-driven medication preferences (see Vansteenwegen et al. 2006).

These effects of DTCA are salient for policymakers tasked with helping professional medical associations meet their goal of preserving the therapeutic orientation of doctor-patient relationships. That goal faces the challenge of increasing commercialization of medical practice at a time when patients are becoming more forthright and assertive in relation to their health care. Thus, the Australian Medical Association's *Code of Ethics* advises that doctors should "Recognize that an established therapeutic relationship between doctor and patient must be respected" (Australian Medical Association 2006, Sect. 1.1.14). Allowing DTCA of prescription pharmaceuticals in countries such as the United States and New Zealand seems to have shifted many doctors' governing conditions from upholding patients' best interests, to meeting patients' brand-specific drug requests even when the request has questionable clinical merit.

Policymakers should be wary of the potential for DTCA to redefine doctor-patient relationships as consumer relationships. In return for being granted a monopoly of expertise on the provision of key goods, doctors are expected to have certain professional character traits. Doctors should be guided by a disposition to serve their patients' best interests and to prioritize patient welfare in their decisions. Doctors make a commitment on joining the medical profession to display a certain kind of professional character, one that, among others, maintains therapeutic rather than commercial relationships with their patients (see Oakley 2014). Even if DTCA were deemed beneficial on the grounds of health outcomes (though the cited evidence makes this unlikely) or perhaps by appeal to freedom of speech, its erosion of the therapeutic doctor-patient relationship is too great a moral cost to pay.

Conclusion

Research in social psychology increasingly shows that advertising persuades subtly, often outside of awareness. Evaluative conditioning is a well-researched implicit persuasive technique that plausibly operates in DTCA. There are good reasons to suppose that evaluative conditioning leads to more positive attitudes toward the advertised drug, and that such attitudes encompass inflated beliefs about drug safety and efficacy. Of grave concern, however, is that the positive imagery deployed to produce such conditioned beliefs bears little substantive relationship with properties of the advertised drug. Given the materiality of these drug properties for people contemplating pharmacological treatment, such an unreliable influence on belief will likely undermine their justification, and antagonize the autonomy of resulting medical decisions.

It is recognized that much recent empirical work in psychology and theorizing in the philosophy of agency raise serious questions about the extent to which conscious deliberation influences ultimate choice. Yet, as outlined, this work may not undermine the case against implicit persuasion in DTCA. Rather, given the augmented role for automatic processing in human choice making, greater efforts should arguably be made to prevent this system from being corrupted. Reduced exposure to implicit techniques such as evaluative conditioning presents as one means by which that end might be realized.

Nor is it sufficient to place the burden of preserving patients' autonomous choices on the physician gatekeeper. While the presence of that learned intermediary may mitigate heteronomous decision-making, strong evidence suggests it does so imperfectly. Physicians are vulnerable to pressure and persuasion from patients motivated, sometimes by wishful thinking, to realize their goal of pharmacological treatment.

On each of these counts, the emerging evidence of implicit persuasion as a driver of advertising effectiveness should counsel caution about DTCA. This should give reassurance to those many countries who outlaw DTCA. By contrast, the United States and New Zealand must widen their regulatory ambit to cover implicit persuasion in DTCA and encourage more research to delineate its effects.

Cross-References

- ▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
- ▶ Ethics of Neuromarketing: Introduction
- ▶ Neuromarketing: What Is It and Is It a Threat to Privacy?
- ▶ Using Neuropharmaceuticals for Cognitive Enhancement: Policy and Regulatory Issues

References

Australian Medical Association. (2006). *AMA code of ethics*. Retrieved August 12, 2013, from https://ama.com.au/codeofethics

Bargh, J., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54*, 462–479.

Benkert, O., Graf-Morgenstern, M., Hillert, A., Sandmann, J., Ehmig, S. C., Weissbecker, H., . . . Sobota, K. (1997). Public opinion on psychotropic drugs: An analysis of the factors influencing acceptance or rejection. *Journal of Nervous and Mental Disease, 185*(3), 151–158.

Biegler, P., & Vargas, P. (2013). Ban the sunset? Nonpropositional content and regulation of pharmaceutical advertising. *The American Journal of Bioethics, 13*(5), 3–13.

Biegler, P. & Vargas, P. *Feeling is believing: Evaluative conditioning and regulation of pharmaceutical advertising* (in preparation).

Biegler, P., Vargas, P., & Oakley, J. *Persuasive effects of non-propositional content in a prescription sleeping pill commercial* (Unpublished study).

Bizer, G. Y., Barden, J. C., & Petty, R. E. (2003). Attitudes. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (Vol. 1, pp. 247–253). Hampshire: Macmillan.

Bortolotti, L. (2011). Does reflection lead to wise choices? *Philosophical Explorations, 14*(3), 297–313.

Bradley, M. M., & Lang, P. J. (2007). *The international affective digitized sounds (2nd edition; IADS-2): Affective ratings of sounds and instruction manual*. Gainesville: University of Florida.

Brass, E. P. (2001). Changing the status of drugs from prescription to over-the-counter availability. *New England Journal of Medicine, 345*(11), 810–816.

Brunstrom, J. M., & Higgs, S. (2002). Exploring evaluative conditioning using a working memory task. *Learning and Motivation, 33*, 433–455.

Carruthers, P. (2005). *Consciousness: Essays from a higher-order perspective*. Oxford: Oxford University Press.

Chartrand, T. L. (2005). The role of conscious awareness in consumer behavior. *Journal of Consumer Psychology, 15*(3), 203–210.

Chartrand, T. L., Huber, J., Shiv, B., & Tanner, R. J. (2008). Nonconscious goals and consumer choice. *Journal of Consumer Research, 35*(2), 189–201.

Christman, J. (1991). Autonomy and personal history. *Canadian Journal of Philosophy, 21*(1), 1–24.

De Houwer, J. (2009). Conditioning as a source of liking: There is nothing simple about it. In M. Wanke (Ed.), *Social psychology of consumer behavior*. New York: Taylor and Francis.

De Houwer, J., Baeyens, F., & Eelen, P. (1994). Verbal evaluative conditioning with undetected US presentations. *Behaviour Research and Therapy, 32*(6), 629–633.

Dijksterhuis, A., & van Olden, Z. (2006). On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology, 42*(5), 627–631.

Ditto, P. H., Jemmott, J. B., 3rd, & Darley, J. M. (1988). Appraising the threat of illness: A mental representational approach. *Health Psychology, 7*(2), 183–201.
Doris, J. M. (2009). Skepticism about persons. *Philosophical Issues, 19*, 57–91.
Drew, T., Vo, M. L., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science, 24*(9), 1848–53.
Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge/New York: Cambridge University Press.
Eifert, G. H., Craill, L., Carey, E., & O'Connor, C. (1988). Affect modification through evaluative conditioning with music. *Behaviour Research and Therapy, 26*(4), 321–330.
Faden, R. R., Beauchamp, T. L., & King, N. M. P. (1986). *A history and theory of informed consent*. New York: Oxford University Press.
Fang, X., Singh, S., & Ahluwalia, R. (2007). An examination of different explanations for the mere exposure effect. *Journal of Consumer Research, 34*, 97–103.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge/New York: Cambridge University Press.
Food and Drug Administration. (2012). *Prescription drug advertising: Questions and answers*. Retrieved April 23, 2013, from http://www.fda.gov/Drugs/ResourcesForYou/Consumers/PrescriptionDrugAdvertising/UCM076768.htm
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20.
Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). "I" values freedom but "We" value relationships: Self-construal priming mirrors cultural differences in judgment'. *Psychological Science, 10*(4), 321–326.
Gilbody, S., Wilson, P., & Watt, I. (2005). Benefits and harms of direct to consumer advertising: A systematic review. *Quality & Safety in Health Care, 14*(4), 246–250.
GlaxoSmithKline. *Commerical for Advair Diskus*. Retrieved November 6, 2013, from https://www.youtube.com/watch?v=4uDtay8Mth8
Goldman, A. I., & Olsson, E. J. (2008). Reliabilism and the value of knowledge. In D. Pritchard, A. Millar, & A. Haddock (Eds.), *Epistemic value*. Oxford: Oxford University Press.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. London: Allen Lane.
Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (The cognitive science of morality: Intuition and diversity, Vol. 2, pp. 181–217). Cambridge, MA: MIT Press.
Hasman, A., & Holm, S. (2006). Direct-to-consumer advertising: Should there be a free market in healthcare information? *Cambridge Quarterly of Healthcare Ethics, 15*, 42–49.
Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136*(3), 390–421.
Hutton, R. J., & Klein, G. (1999). Expert decision making. *Systems Engineering, 2*(1), 32–45.
Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning: The "how" question. *Advances in Experimental Social Psychology, 43*, 205–255.
Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.
Kahneman, D., Ritov, I., & Schkade, D. (2000). Economic preferences or attitude expressions? An analysis of dollar responses to public issues. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames*. Cambridge: Cambridge University Press.
Kravitz, R. L., Epstein, R. M., Feldman, M. D., Franz, C. E., Azari, R., Wilkes, M. S., . . . Franks, P. (2005). Influence of patients' requests for direct-to-consumer advertised antidepressants:

A randomized controlled trial. *Journal of the American Medical Association, 293*(16), 1995–2002.

Krosnick, J. A., Jussim, L. J., & Lynn, A. R. (1992). Subliminal conditioning of attitudes. *Personality and Social Psychology Bulletin, 18*(2), 152–162.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53*(4), 636–647.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.

Kuran, T., & Sunstein, C. R. (1999). Availability cascades and risk regulation. *Stanford Law Review, 5*, 683–761.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (Technical report A-8). Gainesville: University of Florida.

Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. New York: Oxford University Press.

Mintzes, B. (2006). *Direct-to-consumer advertising of prescription drugs in Canada: What are the public health implications*? Retrieved April 23, 2013, from http://www.haiweb.org/15022006/DTCA%20of%20Prescription%20Drugs%20in%20Canada.pdf

Mintzes, B. (2009). Should Canada allow direct-to-consumer advertising of prescription drugs?: No. *Canadian Family Physician, 55*(2), 131–133.

Mintzes, B., Barer, M. L., Kravitz, R. L., Kazanjian, A., Bassett, K., Lexchin, J., & Marion, S. A. (2002). Influence of direct to consumer pharmaceutical advertising and patients' requests on prescribing decisions: Two site cross sectional survey. *British Medical Journal, 324*(7332), 278–279.

Mintzes, B., Barer, M. L., Kravitz, R. L., Bassett, K., Lexchin, J., Kazanjian, A., & Marion, S. A. (2003). How does direct-to-consumer advertising (DTCA) affect prescribing? A survey in primary care environments with and without legal DTCA. *Canadian Medical Association Journal, 169*(5), 405–412.

Murray, E., Lo, B., Pollack, L., Donelan, K., & Lee, K. (2003). Direct-to-consumer advertising: physicians' views of its effects on quality of care and the doctor-patient relationship. *Journal of the American Board of Family Medicine, 16*(6), 513–524.

Murray, E., Lo, B., Pollack, L., Donelan, K., & Lee, K. (2004). Direct-to-consumer advertising: Public perceptions of its effects on health behaviors, health care, and the doctor-patient relationship. *Journal of the American Board of Family Medicine, 17*(1), 6–18.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–258.

Oakley, J. (2012, June). *Virtue ethics, direct-to-consumer pharmaceutical advertising, and doctor-patient relationships*. Paper presented at the 11th World Congress of Bioethics, Rotterdam.

Oakley, J. (2014). A virtue ethics analysis of disclosure requirements and financial incentives as responses to conflicts of interest in physician prescribing. In A. Akabayashi (Ed.), *The future of bioethics: International dialogues*. Oxford: Oxford University Press.

Pfizer. (2010). *Lipitor medication 2010 commercial*. Retrieved March 28, 2013, from http://www.youtube.com/watch?v=ogyC9rEjxDM

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(1), 130–144.

Priest, A. (2007). CanWest set to challenge ban on DTCA. *Canadian Medical Association Journal, 176*(1), 19–20.

Rogers v Whitaker. (1992). 67 Australian Law Journal Reports 47.

Rozin, P., Wrzesniewski, A., & Byrnes, D. (1998). The elusiveness of evaluative conditioning. *Learning and Motivation, 29*, 397–415.

Schachtman, T. R., Walker, J., & Fowler, S. (2011). Effects of conditioning in advertising. In T. R. Schachtman & S. Reilly (Eds.), *Associative learning and conditioning theory: Human and non-human applications*. New York: Oxford University Press.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*(3), 513.

Silversides, A. (2001). Direct-to-consumer prescription drug ads getting bolder. *Canadian Medical Association Journal, 165*(4), 462.

Smith, P. W., Feinberg, R. A., & Burns, D. J. (1998). An examination of classical conditioning principles in an ecologically valid advertising context. *Journal of Marketing Theory and Practice, 6*(1), 63–72.

Spurgeon, D. (1999). Doctors feel pressurised by direct to consumer advertising. *British Medical Journal, 319*(7221), 1321.

Sweldens, S., Van Osselaer, S. M. J., & Janiszewski, C. (2010). Evaluative conditioning procedures and the resilience of conditioned brand attitudes. *Journal of Consumer Research, 37*, 473–489.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. New York: Penguin.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 455–458.

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the name-race implicit association test measure racial prejudice? *Experimental Psychology, 58*(4), 271–277.

Vansteenwegen, D., Francken, G., Vervliet, B., De Clercq, A., & Eelen, P. (2006). Resistance to extinction in evaluative conditioning. *Journal of Experimental Psychology. Animal Behavior Processes, 32*(1), 71–79.

Vargas, P. T. (2008). Implicit consumer cognition. In C. P. Haugtvedt, P. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology*. New York: Lawrence Erlbaum Associates.

Walther, E. (2002). Guilty by mere association: Evaluative conditioning and the spreading attitude effect. *Journal of Personality and Social Psychology, 82*(6), 919–934.

Wilson, T., & Schooler, J. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*(2), 181–192.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101–126.

Young, R. (1986). *Personal autonomy: Beyond negative and positive liberty*. London: Croom Helm.

Yuen, K. S. L., & Lee, T. M. C. (2003). Could mood state affect risk-taking decisions? *Journal of Affective Disorders, 75*, 11–18.

Zajonc, R. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2), 1–27.

Section XXII

Developmental Neuroethics

Developmental Neuroethics

105

Martha J. Farah

Contents

Abstract

This introduction outlines the six chapters in the section "Development Neuroethics".

The neurotechnologies, treatments and enhancements of earlier Handbook sections are applicable to people of all ages. Few of the ethical dilemmas of research and clinical practice are specific to subjects or patients at a particular stage of life. Philosophical problems of mind, brain and identity are no more or less puzzling where children are concerned. Why, then, devote a separate section of a neuroethics handbook to developmental neuroethics?

One reason is that neuroethical issues that have been well explored in relation to adults often acquire new twists when the people involved are children. For example, where psychiatric treatment and enhancement are concerned, children's interests are generally entrusted to parents and other adults. This introduces new opportunities for conflicts of interest that we do not encounter with adults, involving for example the desire that parents and teachers may have for more peaceful homes and classrooms (▶ Chap. 106, "Neuroethical Issues in the Diagnosis and Treatment of Children with Mood and Behavioral Disturbances" by Johnston & Parens).

Neuroethical dilemmas may also play out differently in adults and children because the empirical facts of adult and child brain function differ. The greater sensitivity of young brains to environmental influence raises the issue of early interventions, when the individual's need for intervention may be less certain but

M.J. Farah
Center for Neuroscience & Society, University of Pennsylvania, Philadelphia, PA, USA
e-mail: mfarah@neuroethics.upenn.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_102,

the likely effectiveness of it would be higher (▶ Chap. 107, "Prediction of Antisocial Behavior"). This same heightened plasticity argues for greater caution when intervening, for example with pharmacology or noninvasive brain stimulation (▶ Chap. 108, "Mind, Brain, and Education: A Discussion of Practical, Conceptual, and Ethical Issues"; ▶ Chap. 106, "Neuroethical Issues in the Diagnosis and Treatment of Children with Mood and Behavioral Disturbances" by Johnston & Parens).

Finally, some neuroethical issues are inherently linked to certain developmental stages. School takes up much of children's waking lives. It is therefore children who stand to be affected most by neuroeducation, and children who will bear the brunt of any premature or misguided educational practices it spawns (▶ Chap. 108, "Mind, Brain, and Education: A Discussion of Practical, Conceptual, and Ethical Issues"). The search for objective measures of cognitive development and neural maturity, to inform policies concerning children and adolescents and to determine to which individuals the policies apply (▶ Chap. 109, "Normal Brain Development and Child/Adolescent Policy"), is by definition linked to the young. Finally, childhood is the period when most psychological traits emerge, including atypical cognitive styles (Baron-Cohen ▶ Chap. 111, "Neuroethics of Neurodiversity"), gender identification (▶ Chap. 110, "Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences" by Fine) and antisocial personality traits (▶ Chap. 107, "Prediction of Antisocial Behavior"). The neuroscience of these important individual differences are therefore rooted in childhood. Although the neuroethical issues raised in these cases are diverse, from biased interpretations of research data (▶ Chap. 110, "Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences" by Fine) to the challenge of distinguishing between difference and deficit (Baron-Cohen, ▶ Chap. 111, "Neuroethics of Neurodiversity"), all exemplify the important work underway in developmental neuroethics.

References

Ansari, D. (2014). Mind, brain and education: A discussion of practical, conceptual and ethical issues. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics*. New York: Springer.

Baron-Cohen, S. (2014). Neuroethics of neurodiversity. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics*. New York: Springer.

Fine, C. (2014). Neuroscience, gender and "development to" and "from": The example of toy preferences. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics*. New York: Springer.

Glenn, A. L., Foquert, F., & Raine, A. (2014). Prediction of antisocial behavior. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics*. New York: Springer.

Johnson, S. B., & Giedd, J. N. (2014). Normal brain development and child/adolescent policy. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics*. New York: Springer.

Johnston, J., & Parens, E. (2014). Neuroethical Issues in the diagnosis and treatment of children with mood and behavioral disturbances. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics*. New York: Springer.

Neuroethical Issues in the Diagnosis and Treatment of Children with Mood and Behavioral Disturbances

106

Josephine Johnston and Erik Parens

Contents

J. Johnston (✉) • E. Parens
The Hastings Center, Garrison, NY, USA
e-mail: johnstonj@thehastingscenter.org; parense@thehastingscenter.org

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_147,

Abstract

The number of children in the United States receiving psychiatric diagnoses and taking psychotropic medications rose significantly from the second half of the twentieth century through to today. Accompanying these increased rates of diagnosis and psychotropic medication use have come sometimes intense debates about whether the increases are appropriate, or whether healthy children are being mislabeled as sick and inappropriately given medications to alter their moods and behaviors. While these debates are in part highly technical, concerning questions in epidemiology and pharmacology, they are also infused with ethical questions about the appropriate goals of medicine, the nature of sickness and health, and the obligations we owe to children and families struggling to flourish. This chapter presents four inter-connected observations about the diagnosis and treatment of mood and behavioral disturbances in children that at least partially explain why this area generates concern and controversy, but that also point to important areas of agreement where progress can be made. These observations include that psychiatry can provide an important approach for understanding and responding to children's mood and behavioral problems provided we remember that it also carries its own complexities and difficulties. Further, forces within and outside psychiatry can influence how diagnoses are made and how treatments are selected, including systemic forces that strongly favor medication over psychosocial treatments, with the result that children too often receive pharmacological treatment only, even when other interventions are supported by evidence, contribute towards long-term flourishing, and reflect families' deepest value commitments.

Introduction

The number of children in the United States receiving psychiatric diagnoses and taking psychotropic medications rose significantly from the second half of the twentieth century through to today. A CDC report issued in 2013 estimated that 6.8 % of children aged 3–17 years living in the United States have a current diagnosis of attention-deficit/hyperactivity disorder (ADHD), with parents of 8.5 % of children reporting that they had ever been told that their child had ADHD (CDC 2013). While ADHD is by far the most widely diagnosed condition in children, parents report a current diagnosis of behavioral or conduct problems for 3.5 % of US children, anxiety for 3.0 %, depression for 2.1 %, and autism spectrum disorders for 1.1 %. As the CDC report also noted, almost 20 years' of surveillance show that these diagnostic rates have increased over time. Increases in diagnostic rates have been accompanied by increases in overall rates of use of psychotropic medications in children (Olfson et al. 2002; Zito et al. 2003), particularly antidepressants (Delate et al. 2004), stimulants (Safer et al. 1996; Habel et al. 2005), and anti-psychotics (Patel et al. 2005; Olfson et al. 2010).

Accompanying these increased rates of diagnosis and psychotropic medication use have come sometimes intense debates about whether the increases are appropriate, or whether healthy children are being mislabeled as sick and inappropriately given medications to alter their moods and behaviors (Timimi and Taylor 2004; Coghill 2004; Biederman 2005; Singh 2004; Conrad and Potter 2000; Olfman 2006; Jensen et al. 2006). While these debates are in one part highly technical, concerning questions in epidemiology and pharmacology, they are also infused with ethical questions about the appropriate goals of medicine, the nature of sickness and health, and the obligations we owe to children and families struggling to flourish.

Our chapter is organized around four inter-connected observations about the diagnosis and treatment of mood and behavioral disturbances in children that at least partially explain why this area generates concern and controversy, but that also point to important areas of agreement where progress can be made.

The First General Observation Is that Society Has an Obligation to Help Children Who Are Suffering from Mood and Emotional Disturbances, But that Deciding How to Help, and Delivering that Help, Can Be Practically and Politically Difficult

Although it is generally accepted that we should help troubled children there is controversy about *how* best to fulfill this obligation. Helping troubled children can involve making changes not – or not only – in the child herself, but also in her environment (her family, her school, her neighborhood, etc.). Children are deeply and tightly embedded in environments (or contexts) that can play a role in the emergence or severity of their mood and behavioral disturbances and, importantly, that can be key to their improvement. However, changing these environments can be practically difficult. For instance, it can involve changing classroom size, providing different kinds of instruction and classrooms, alleviating poverty, preventing home and neighborhood violence, changing parenting practices, and improving nutrition. It can also be politically difficult. For instance changing parenting practices, may be considered an intrusion on a matter of private choice. Helping parents, financially and in other ways, may raise questions about the just distribution of collective resources.

The Second Observation Is that Psychiatry Provides an Important and Powerful Approach for Understanding and Responding to Children's Moods and Behaviors, But that This Approach Carries Its Own Complexities and Difficulties

When children's mood and behavioral disturbances are understood as possible symptoms of a psychiatric disorder, the diagnostic categories and treatment tools of psychiatry can be used. In the US, this approach involves diagnosis with the American Psychiatric Association's *Diagnostic and Statistical Manual* (DSM).

The diagnostic categories in the DSM have been created over time by committees of experts, drawing on clinical experience and published research. The categories in the DSM are generally not based on an understanding of the pathophysiology of the clusters of symptoms they name, and most DSM diagnoses are made without the use of physiological tests (Nature Editorial 2010). Instead, a psychiatric diagnosis is today a judgment about whether a child meets the diagnostic criteria. The judgment is based on: the clinician's interpretation of the disorder's diagnostic criteria; the clinician's training and clinical experience; the clinician's observations of the child during the appointment; parents' and possibly teachers' and school psychologists' reports of the child's moods and behaviors; and, often, the results of a diagnostic instrument like a symptom checklist or structured interview.

Acknowledging the role of judgment, and the possibility of disagreements among clinicians, does not imply that childhood psychiatric diagnoses are not real. The clusters of signs and symptoms described in the DSM can cause real suffering in children (Hinshaw 1992; Barkley 2006) and real costs to families, the health care system, the education system, the juvenile justice system, and employers (through parental work loss) (Welham et al. 2007). Instead, it reminds us that psychiatric diagnoses are tools for thinking about the very real, varied, and sometimes deeply difficult lived experience of adults and children. Wielded thoughtfully, with awareness of their socially constructed nature, those categories can help to identify children who can benefit from intervention.

Several facts about the DSM approach help to explain why psychiatric diagnoses in general can generate debate and why that debate can be particularly acute when they are applied to children.

Children with Different Symptoms Can Receive the Same Diagnosis

For example, according to DSM-5, the essential feature of ADHD is "a persistent pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development" (APA 2013). To receive the ADHD diagnosis, children must exhibit at least five or six (depending on the child's age) of the 18 core symptoms listed in the manual. The symptoms are divided into two major behavioral domains: inattention and impulsivity-hyperactivity. Among the nine symptoms of inattention: often making careless mistakes, often having difficulty sustaining attention in play or other activities, and often not seeming to listen when spoken to directly. A child exhibits some of the nine symptoms of hyperactivity-impulsivity if the child often fidgets or squirms, often cannot stay seated, blurts out, and has difficulty awaiting a turn. Different children can exhibit a different cluster of these 18 behaviors, but receive the same diagnosis.

Children with Some of the Same Symptoms Can Receive Different Diagnoses

For example, according to DSM-5, to receive a diagnosis of classic or full-blown bipolar disorder (bipolar I), the individual must experience a manic episode, which is

"a distinct period of abnormally and persistently elevated, expansive, or irritable mood" lasting for at least 1 week. If the patient's mood is elevated or expansive she must exhibit at least three of the following seven symptoms: (1) grandiosity, (2) decreased need for sleep, (3) pressure to keep talking, (4) flight of ideas and racing thoughts, (5) distractibility, (6) increased goal-directed activity and psychomotor agitation, or (7) excessive involvement in pleasurable activities that have a high potential for painful consequences. If the patient presents with irritability, she must exhibit at least four of those seven symptoms. At a minimum, three of the symptoms used to diagnose bipolar disorder are very similar to those used to diagnose ADHD: pressure to keep talking, psychomotor agitation, and distractibility. If one adds into the mix the symptoms of oppositional defiant disorder (ODD), which is frequently characterized by irritable mood, it can be difficult to determine whether bipolar disorder, ADHD, or ODD – and now Disruptive Mood Dysregulation Disorder (DMDD) – is the best-fitting diagnosis. In practice, one result of this overlap is that children showing a mix of symptoms often receive more than one diagnosis (and are treated with more than one medication). Another result is that children who exhibit some of the same symptoms may receive different diagnoses.

Symptoms of the Same Disorder Can Look Different in Children and Adults

Unlike its predecessor, DSM-5 does not contain a special section of disorders usually first diagnosed in infancy, childhood, or adolescence, but instead integrates diagnoses most applicable to children throughout the manual by placing them at the beginning of diagnostic chapters (e.g., the first disorder listed in the Anxiety Disorders chapter is Separation Anxiety Disorder). Despite this integration, clinicians may still diagnose children with disorders usually reserved for adults (or vice versa), which may require adapting the diagnostic criteria. Two examples of this adaptation process can be seen in depression and bipolar disorder. Before the 1970s, clinicians theorized that, while children could experience transient sadness, they were not sufficiently emotionally developed to experience clinical depression (Abela and Hankin 2008). By the 1980s, researchers argued that children can experience depression, but that depressive symptoms can take slightly different forms in adults and children. For example, while adults may experience depressed mood and significant loss of interest in activities, small children may be more inclined to show particularly severe separation anxiety, while restlessness, sulkiness, and withdrawal from social activities might be more pronounced in adolescents (Kashani et al. 1981). Today, the idea that children can experience depression and that their symptoms may be different from those seen in adults is fairly uncontroversial within psychiatry, even if there remains some debate about how best to treat it (March and Vitiello 2009).

The story of the diagnosis of bipolar disorder in children is quite different. While it is widely agreed within pediatric psychiatry that some rare children exhibit discrete episodes of mania and meet full DSM criteria for bipolar disorder, much of the controversy in the United States over the past decade

has been rooted in disagreements about whether it can look *very* different in children and adults. Beginning in 1995, some researchers began to argue that *chronic* irritability (or raging) was a symptom of mania in children, even though in adults clinicians look for *distinct* episodes of "abnormally and persistently elevated, expansive or irritable mood" (APA 2013). That argument was highly contested, although it is not theoretically implausible. A subset of adults with bipolar say that their symptoms went unnoticed when they were children and children's bodies are developing and are different from adults', so it is conceivable that prodromal symptoms of a disorder, including bipolar disorder, or symptoms of the full-blown disorder itself, could look quite different in children and adults. However, in the case of bipolar disorder, which began to be much more frequently diagnosed in children during the 2000s, some researchers argued that the symptoms at issue, in particular chronic irritability, are best understood as markers of a different disorder altogether. In 2003, one team began using the term "severe mood dysregulation" to describe these children (Leibenluft et al. 2003), and in 2013 DSM-5 introduced the new diagnosis of "Disruptive Mood Dysregulation Disorder (DMDD)" for children exhibiting severe recurrent temper outbursts in response to common stressors (APA 2013). It is not yet known the extent to which DMDD will replace previous diagnoses of bipolar disorder in children.

Careful Diagnosis Requires Both Identification of Symptoms and Evaluation of Impairment

DSM-5 is clear that the presence of symptoms alone seldom warrants a diagnosis; a diagnosis is warranted when symptoms create significant impairment. Some impairment might be inferred from the fact that parents make appointments with health professionals, but impairment assessments are unfortunately not always included in diagnostic work-ups. When they are included, diagnostic rates are lower. In one study, researchers assessing a sample of children for serious emotional disturbances found prevalence rates of between 4 % and 8 %, depending on which of three different impairment measures was used, and a prevalence rate of 20 % when impairment was ignored (Costello et al. 1996). Reimbursement systems, which require a DSM diagnosis, may indirectly encourage clinicians to record a diagnosis even when the severity criteria are not fully met, in order to justify the provision of services.

The Diagnostic System Does Not Encourage Assessment of the Child's Context

In the context of depression, Allan Horwitz and Jerome Wakefield have argued that "the basic flaw" of the DSM approach is that it "*fails to take into account the context of the symptoms*" (Horwitz and Wakefield 2007). Writing about

depression as defined in DSM-IV, Horwitz and Wakefield noted that only intense sadness in response to the death of a loved one was listed as a context-specific reaction that should not count as a symptom, whereas myriad other sorts of normal human problems that could also trigger intense sadness – from the lack of strong, meaningful attachments to job loss (in adults) to being bullied or neglected (in children) – went unmentioned. As a result, Horwitz and Wakefield argued, people who are intensely but appropriately sad due to life events or circumstances can mistakenly receive a diagnosis of depression. (They are thinking primarily of adults, but the same analysis applies to children.) Their complaint would likely be even stronger now that DSM-5 has removed the bereavement exclusion from the definition of depression (Wakefield 2013).

Robert Spitzer, the head of the American Psychiatric Association's DSM-III task force, has similarly noted that the definition of mental disorder offered in the *introduction* to the DSM clearly states that mental disorder involves dysfunction or impairment that is not an expectable or proportionate response to a common human problem or stressor, but the diagnostic criteria used in the *body* of the manual – the part that clinicians usually consult – rarely mention the need to consider contextual explanations for symptoms. According to Spitzer, this arrangement allows "normal responses to stressors to be characterized as symptoms of disorder" (Horwitz and Wakefield 2007). By failing to discuss contextual explanations for problematic moods and behaviors, the manual can seem to suggest that context is irrelevant to diagnosis and treatment decisions. This apparent failure is particularly important when dealing with children. If a child's moods and behaviors are an adaptive or appropriate response to her adverse, traumatic, or otherwise difficult context, it could be a serious mistake to treat the child but fail to make changes to her environment. But it is also important to remember that a contextual explanation does not by itself indicate that the child is not suffering from a mental disorder. Just as a child whose fever results from drinking unclean water needs both a fever medication *and* an improved water supply, so an abused child suffering posttraumatic stress disorder (or another condition likely linked to their abuse) may be helped both by treatment (pharmacological and/or psychosocial) and changes to her environment.

Symptoms and Impairment Are Dimensional, and Children Are Developing Organisms

The introduction to DSM-5 acknowledges that psychiatric diagnoses refer to phenomena that are *dimensional*, but the body of the text uses *categories* to name them. This tension is acknowledged by the manual's authors, although in the end they conclude that it is "premature scientifically to propose alternative definitions for most disorders" (APA 2013). When the DSM-5 authors use "dimensional" in the introduction, they refer to the fact that symptoms appear on a continuum of

expression or intensity, and that so, too, can disorders. Individuals who, for example, exhibit a single symptom such as excessive worrying can do so to different degrees. And individuals who exhibit a cluster of symptoms indicative of anxiety disorder can also do so to different degrees, which can produce different degrees of impairment.

Determining whether a given child's moods and behaviors are intense enough to be labeled disordered is further complicated by the fact that, as still-developing organisms, their moods and behaviors can be very different from those we see in adults and can vary greatly depending on the age of the child (it may be normal for a 4-year-old child to talk with an imaginary friend, but not for a 14-year-old or an adult) (McClellan 2005). Indeed, the experiences of children who do and do not live "under the description of" a psychiatric disorder, as the anthropologist Emily Martin wrote in 2007, are not always as radically different as the categorical labels can seem to suggest. This dimensionality is not unique to children, or to psychiatry. There is also a continuum between adults who do and do not warrant a diagnosis of, for example, hypertension. But because our moods and behaviors are closer to our sense of identity than a trait like blood pressure, and because recognizing these moods and behaviors as symptoms of a disorder requires greater observer interpretation than reading blood pressure results, our values play a bigger role in determining where to draw the line on the anxiety continuum than on the blood pressure continuum.

Diagnostic Categories Create a "Zone of Ambiguity"

Because of this dimensionality, there will be significant agreement that some children are on one end of a continuum and need help in changing their impairing moods and behaviors, and that other children are closer to the middle of that continuum and deserve to be affirmed in their atypical-but-not-impairing ways of being. Or, in more colloquial parlance, there will be ready agreement that some atypical children are sick and that other atypical children are healthy. However, there will be a sizable "zone of ambiguity" between those uncontested regions of the continuum, in which reasonable people may well disagree about where to draw the line between normal and disordered, about whether or not a given child is suffering from a disorder. This disagreement may result from differences in knowledge about, or experience of, children and their moods and behaviors, but it can also result from differences in the underlying value commitments of these reasonable people. Some may have an expansive conception of disordered behavior, and others may have an expansive conception of normal variation. Acknowledging the existence of this zone of ambiguity and the role that value commitments play in diagnostic decisions made within this zone does not undermine the seriousness of the problems that families and children experience (although it can sound that way to some who deal with these problems day to day (Resko 2011)).

The Third General Observation About the Ethical Issues Surrounding the Treatment of Troubled Children, Is that Values Play an Ineliminable Role in the Creation of Diagnostic Systems, the Diagnosis of Individual Children, and in Treatment Decisions

Again, recognizing that judgment and therefore values play a role in the creation of diagnostic categories does not diminish the potential harmfulness of the moods and behaviors at issue, nor imply that those categories and treatments are arbitrary or useless. Instead, it forces us to be clear about the extent to which medicine, including psychiatry, is socially constructed.

People can, as a result of different value commitments, hold different views about how narrow or broad the goals of pediatric psychiatry should be. When a child's symptoms land her in the zone of ambiguity, those differences can affect diagnosis. Psychiatry is not unique in harboring disagreements about how narrow or broad our conceptions of illness and health should be – nor about how cautious or aggressive our treatment approaches should be. Some observers are untroubled by the tendency of medicine in general – and psychiatry in particular – to treat problems that seem to have their proximate cause in educational, social, or cultural mores rather than in pathophysiological dysfunctions. Such observers can argue that, insofar as the goal of medicine and psychiatry is to promote the well-being of persons, and insofar as what counts as well-being always depends on functioning in a particular time and place, there is no reason to be alarmed if psychiatrists aim to help people to function – or even to excel – in this particular time and place (Greely et al. 2008). Other observers are alarmed by this tendency. They suggest that these expanded goals and lowered diagnostic thresholds pose risks to individuals and society (Conrad 2007). Critics of 'medicalization,' as it has been called by Peter Conrad, are concerned that it is too often fueled not by the needs of patients but by drug companies, which profit by creating or expanding disorders for which they then market medication treatments, even where the medications have limited efficacy and carry the risk of serious side effects (Conrad 2007; Sadler et al. 2009; Sadler 2007). Others, like William Carey, are concerned that we are losing touch with what is normal for children or that the definition of normal in children is getting narrower and narrower for the convenience of adults and not the best interest of children (Carey 2011). Conrad, Carey, and others demand that we recognize that a wide range of human temperaments and behaviors are compatible with a healthy human life (Carey and McDevitt 1995).

Whether one is more or less distressed by medicalization of children's moods and behaviors can partly depend upon the extent to which one emphasizes one of two deep obligations that parents must constantly balance (Parens 2006). On the one hand, parents have an obligation to let their children unfold in their own ways, to affirm them as individuals and to let them be who they are. The violin-loving father who pushes his football-loving son to play the violin fails to accept his son and affirm his son's pursuit of what seems good to him. On the other hand, parents

have an obligation to shape their children through discipline, education, and adherence to traditions. A parent who lets her child stay home all day every day to play video games violates her obligation to shape her child.

Though both obligations are fundamentally important, it is inevitable that in particular situations some parents will emphasize one over the other. Parents who emphasize their obligation to shape their children may be fairly quick to see intervention in the zone of ambiguity as just one more instance of fulfilling that obligation – even though they accept that they also have an obligation to let their children unfold in their own way. Other parents will be more inclined to let their children unfold in their own ways and will therefore be reluctant to see their children's moods and behaviors as potentially "disordered" and in need of psychiatric assessment.

The situation can be similar when choosing which *means* to use to treat a child. Few dispute that medication should play a role in the treatment of children with classic bipolar disorder, and few dispute that behavioral therapies should play a role in the treatment of children with depression (Whittington et al. 2004). Yet as we found in the case of ADHD, disagreement can occur when the data on the efficacy of various treatments is unclear (Fabiano et al. 2009; UK NICE 2008; Schlander 2008). In the face of this complexity and disagreement, parents and clinicians may prefer one or the other means of treatment not simply because of what clinical trials have shown about its safety and effectiveness, but because it best fits their preexisting value commitments.

For example, medications tend to emphasize the value of *efficiency* insofar as they are often quicker acting, cheaper in the short term, and require less time to administer than psychosocial treatments. They can quickly improve a child's symptoms so that she can return home from hospital, return to school, or return to her regular activities. Behavioral interventions, on the other hand, tend to emphasize the value of *engagement*, by requiring parents, peers, teachers, or therapists to work with the child and with his environment (Parens 1998). Because behavioral interventions seem to locate the "problem" in the interaction between the child and her home, school, and social context rather than in her body, they can prompt us to notice the importance of the child's environment and take steps to improve it, and they may help the child learn to think of herself as a moral agent, as someone who can learn how to change. While some parents and clinicians will favor the value of efficiency and others will favor the value of engagement, most will acknowledge the importance of both values, just as they appreciate both the obligation to shape children and the obligation to let them unfold in their own ways.

Recognizing that some disagreements about how to define mental disorders, and how to diagnose or treat a given child, can arise because reasonable people emphasize different but equally respectable values in no way minimizes the enormous social and economic pressures bearing on families to emphasize some value commitments rather than others. Nor does it in any way minimize the need to distinguish between reasonable disagreements and mistakes.

In a perfect world, the debate about diagnosing and medicating children would be about how best to balance these different value commitments. But too often in the United States, diagnostic and treatment decisions are driven and constrained by the broader culture and the institutions and systems in which parents, children, and clinicians must operate.

The Fourth and Final Observation Concerning the Diagnosis and Treatment of Childhood Mood and Behavioral Disturbances Is that a Number of Social and Economic Forces Shape Diagnosis and Constrain Treatment Choice

In particular, these forces systematically favor pharmacological over psycho-social interventions. In part because of this, many children do not receive careful diagnoses, evidence-based treatments are often not available, and promising changes to children's environments are not made. Many aspects of current biomedical research funding and healthcare economics shape the diagnosis and treatment of troubled children. Several of these will be reviewed here (see Parens and Johnston 2011 for a more complete discussion).

To begin with, the treatment marketplace is dominated by psychotropic medication, despite evidence supporting the safety and effectiveness of some psychosocial treatments for particular disorders. One reason for this dominance is spending on drug development and testing far exceeds spending to develop and test psychosocial treatments (DiMasi et al. 2003; Adams and Brantner 2006). In addition, psychotropic drug treatments are aggressively marketed to practitioners and patients (Gagnon and Lexchin 2008). This is true despite ongoing concerns about the drug approval process and, therefore, about drug safety and effectiveness (Okie 2005). These concerns are not limited to drugs used to treat children diagnosed with mental disorders, but given underlying worries about the impact of medication on the developing brains and bodies of children and the heightened ethical obligations that physicians and parents have to minors for whom – or with whom – they are making treatment decisions, the concerns take on a particular urgency in this context (Zito et al. 2008).

Changes that would begin to redress the imbalance between investments in the development of new pharmacological compared with psychosocial treatments include sustained or increased government and philanthropic funding of basic research likely to lead to new psychosocial interventions, and of clinical research to test their effectiveness once developed. Once new, evidence-based psychosocial treatments are available, funds will be required to market these treatments and to train practitioners to use them effectively. Changes that would begin to improve the information available about medication treatments as they are actually used in the community include enabling the FDA to require robust post-marketing registries on selected medications that are used in children.

Several features of U.S. health care increase the likelihood that diagnostic mistakes will occur and that psychotropic medications alone will be the default treatment for children's mood and behavioral disturbances. Several of these features are causes for concern in themselves because in addition to limiting clinicians and parents' choices, they suggest that children are not receiving recommended care (Druss 2011).

In general, visits to medical practitioners are very brief. Although one study showed that pediatricians spent an average of between 5 and nearly 7 min longer with patients when behavioral health concerns were raised than when they were not (Cooper et al. 2006), visits including behavioral health concerns are still likely to last less than 20 min. It is extremely difficult in such a limited time for practitioners to undertake careful mental health diagnoses; reassess these diagnoses; discuss and reassess medication treatments; or provide and monitor psychosocial interventions. Not only are these visits of short duration, but they are less frequent than is necessary for optimal treatment management.

With respect to providers, the system is fragmented among primary care physicians, hospitals, and various other mental health care providers, as well as other systems that care for children, including child protective services, juvenile justice, and schools (Foy and Perrin 2010). Practitioners and parents seeking psychosocial interventions have limited ability to identify services, judge their quality, or assess the expertise of individual practitioners. Primary care providers have limited ability to monitor the costs and outcomes of any psychosocial interventions they recommend. When psychosocial services are identified, long waiting lists often delay access, and high rates of staff turnover among mental health providers can disrupt continuity of care. Families who are committed to psychosocial treatments may be left to identify, access, and navigate them alone (Olfson and Marcus 2010).

Where mental health care is funded through private insurance, coverage for psychosocial treatments is often more limited than for medication treatments (Teich and Buck 2007), despite new legislation (Druss 2011). Under managed care plans, medication treatments for emotional and behavioral disorders do not count as behavioral health care costs, but instead fall under the plan's general prescription drug coverage (Frank et al. 2005). Behavioral health care management organizations, therefore, have an incentive to reduce utilization of psychosocial treatments (and hospitalization), but they are unaffected by the use of psychotropic medications. They may further restrict reimbursement for psychosocial interventions by requiring the presence of the patient at each treatment session, which means that they do not cover parent training, for example, which is known to be effective but does not require the presence of the child.

Conclusion and Future Directions

Some of the complexities associated with the current approach to diagnosing emotional and behavioral disturbances in children have been described. Most of the diagnoses articulated in the DSM were based on observation of symptoms in

adults, but symptoms of what psychiatrists consider to be the same disorder may look different in adults and children. Also, the DSM's categories capture heterogeneous phenomena, and they overlap; further, because symptoms and impairments are expressed along continua, there are no bright lines between healthy children and those who warrant diagnoses.

Informed, trained, caring people will thus sometimes have reasonable disagreements about where to set diagnostic thresholds and about whether a mildly affected child – a child in the "zone of ambiguity" – would benefit from a diagnosis. These disagreements can occur when people have different value commitments or just give different emphases to shared value commitments (regarding, for example, the goals of psychiatry or the goals of parenting). Such value differences or emphases can play out in the context of treatment decisions as well.

Nevertheless, some conclusions are widely accepted. For one thing, there is agreement that children can indeed have serious psychiatric disorders and that medications can be an essential part of appropriate treatment plans. For another, no matter how important it is to tolerate reasonable disagreements, it is essential to avoid the sorts of mistakes that involve patent over-diagnosis, misdiagnosis, and under-diagnosis, which result in many children not receiving the care they need. These mistakes are facilitated by systemic forces that bear on clinicians and families and restrict the time available for careful diagnoses. Specifically, these forces can make it tempting to base a diagnosis on the presence of symptoms alone, as opposed to doing the sort of careful evaluation that can determine whether those symptoms impair the child and what role, if any, the child's context might play in causing their distress. Those same systemic forces strongly favor medication treatments over psychosocial ones, so that children too often receive pharmacological treatment only, even when other treatment plans are supported by evidence, contribute towards long-term flourishing, and reflect their or their family's deepest value commitments.

Acknowledgments This chapter is based on Erik Parens and Josephine Johnston (2008): www.capmh.com/content/2/1/5, and Erik Parens and Josephine Johnston (2011). This research was funded by grant U13 MH78722 of the National Institute of Mental Health and by Dr. Eve Hart Rice and Dr. Timothy D. Mattison.

Cross-References

- ▶ Developmental Neuroethics
- ▶ Ethics in Psychiatry
- ▶ Ethics of Pharmacological Mood Enhancement
- ▶ Normal Brain Development and Child/Adolescent Policy
- ▶ The Morality of Moral Neuroenhancement
- ▶ What Is Normal? A Historical Survey and Neuroanthropological Perspective

References

Abela, J. R. Z., & Hankin, B. L. (Eds.). (2008). *Handbook of depression in children and adolescents*. New York: Guilford Press.

Adams, C. P., & Brantner, V. V. (2006). Estimating the cost of new drug development: Is it really $802 million? *Health Affairs, 25*(2), 420–428.

American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders, (DSM-5)*. Arlington: American Psychiatric Publishing.

Barkley, R. A. (2006). Attention-deficit hyperactivity disorder. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 63–112). New York: Guilford Press.

Biederman, J. (2005). Attention-deficit/hyperactivity disorder: A selective overview. *Biological Psychiatry, 57*, 1215–1220.

Carey, W. B. (2011). Primary care physicians need a better understanding of temperamental variation. *Hastings Center Report, 41*(2), S14.

Carey, W. B., & McDevitt, S. C. (1995). *Coping with children's temperament: A guide for professionals*. New York: Basic Books.

Centers for Disease Control and Prevention (CDC). (2013). Mental health surveillance among children – United States, 2005–2011. *Morbidity and Mortality Weekly Report, 62*(Suppl. 2), 1–35.

Coghill, D. (2004). Use of stimulants for attention deficit hyperactivity disorder. *British Medical Journal, 329*, 907–908.

Conrad, P. (2007). *The medicalization of society: On the transformation of human conditions into treatable disorders*. Baltimore: Johns Hopkins University Press.

Conrad, P., & Potter, D. (2000). From hyperactive children to ADHD adults: Observations on the expansion of medical categories. *Social Problems, 47*, 559–582.

Cooper, S., Valleley, R. J., Polaha, J., Begeny, J., & Evans, J. H. (2006). Running out of time: Physician management of behavioral health concerns in rural pediatric primary care. *Pediatrics, 118*(1), e132–e138.

Costello, E. J., Angold, A., Burns, B. J., Erkanli, A., Stangl, D. K., & Tweed, D. L. (1996). The Great Smoky Mountains Study of Youth: Functional impairment and serious emotional disturbance. *Archives of General Psychiatry, 53*(12), 1137.

Delate, T., Gelenberg, A. J., Simmons, V. A., & Motheral, B. R. (2004). Trends in the use of antidepressants in a national sample of commercially insured pediatric patients, 1998 to 2002. *Psychiatric Services, 55*, 387–391.

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics, 22*(2), 151–185.

Druss, B. G. (2011). The changing face of US mental health care. *FOCUS: The Journal of Lifelong Learning in Psychiatry, 9*(2), 221–222.

Fabiano, G. A., Pelham, W. E., Jr., Coles, E. K., Gnagy, E. M., Chronis-Tuscano, A., & O'Connor, B. C. (2009). A meta-analysis of behavioral treatments for attention-deficit/hyperactivity disorder. *Clinical Psychology Review, 29*(2), 129–140.

Foy, J. M., & Perrin, J. (2010). Enhancing pediatric mental health care: Strategies for preparing a community. *Pediatrics, 125*(Suppl. 3), S75–S86.

Frank, R. G., Conti, R. M., & Goldman, H. H. (2005). Mental health policy and psychotropic drugs. *Milbank Quarterly, 83*(2), 271–298.

Gagnon, M. A., & Lexchin, J. (2008). The cost of pushing pills: A new estimate of pharmaceutical promotion expenditures in the United States. *PLoS Medicine, 5*(1), e1.

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456*(7223), 702–705.

Habel, L. A., Schaefer, C. A., Levine, P., Bhat, A. K., & Elliott, G. (2005). Treatment with stimulants among youths in a large California health plan. *Journal of Child and Adolescent Psychopharmacology, 15*, 62–67.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111*, 127–155.

Horwitz, A. V., & Wakefield, J. C. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. New York: Oxford University Press.

Jensen, P., Knapp, P., & Mrazek, D. (2006). *Toward a new diagnostic system for child psychopathology: Moving beyond DSM*. New York/London: The Guilford Press.

Kashani, J. H., Husain, A., Shekim, W. O., Hodges, K. K., Cytryn, L., & McKnew, D. H. (1981). Current perspectives on childhood depression: An overview. *American Journal of Psychiatry, 138*(2), 143–153.

Leibenluft, E., Charney, D. S., Towbin, K. E., Bhangoo, R. K., & Pine, D. S. (2003). Defining clinical phenotypes of juvenile mania. *American Journal of Psychiatry, 160*(3), 430–437.

March, J., & Vitiello, B. (2009). Clinical messages from the treatment for adolescents with depression study (TADS). *American Journal of Psychiatry, 166*(10), 1118–1123.

Martin, E. (2007). *Bipolar expeditions: Mania and depression in American culture*. Princeton: Princeton University Press.

McClellan, J. (2005). Commentary: Treatment guidelines for child and adolescent bipolar disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*, 236–239.

Nature Editorial. (2010). A decade for psychiatric disorders. *Nature, 463*, 9.

Okie, S. (2005). Safety in numbers – Monitoring risk in approved drugs. *New England Journal of Medicine, 352*(12), 1173–1176.

Olfman, S. (Ed.). (2006). *No child left different*. Westport: Praeger Publishers.

Olfson, M., & Marcus, S. C. (2010). National trends in outpatient psychotherapy. *American Journal of Psychiatry, 167*(12), 1456–1463.

Olfson, M., Marcus, S. C., Weissman, M. M., & Jensen, P. S. (2002). National trends in the use of psychotropic medications by children. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*, 514–521.

Olfson, M., Crystal, S., Huang, C., & Gerhard, T. (2010). Trends in antipsychotic drug use by very young, privately insured children. *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(1), 13–23.

Parens, E. (1998). Is better always good? In E. Parens (Ed.), *Enhancing human traits: Ethical and social implications* (pp. 1–28). Washington, DC: Georgetown University Press.

Parens, E. (Ed.). (2006). *Surgically shaping children: Technology, ethics, and the pursuit of normality*. Baltimore: Johns Hopkins University Press.

Parens, E., & Johnston, J. (2008). Understanding the agreements and controversies surrounding childhood psychopharmacology. *Child and Adolescent Psychiatry and Mental Health, 2*, 5.

Parens, E., & Johnston, J. (2009). Facts, values, and Attention-Deficit Hyperactivity Disorder (ADHD): An update on the controversies. *Child and Adolescent Psychiatry and Mental Health, 3*, 1.

Parens, E., & Johnston, J. (2010). Controversies concerning the diagnosis and treatment of bipolar disorder in children. *Child and Adolescent Psychiatry and Mental Health, 4*, 9.

Parens, E., & Johnston, J. (2011). Troubled children: Diagnosing, treating, and attending to context. *Hastings Center Report, 41*(2), S4–S31.

Parens, E., Johnston, J., & Carlson, G. A. (2010). Pediatric mental health care dysfunction disorder? *The New England journal of medicine*.

Patel, N. C., Crismon, M. L., Hoagwood, K., Johnsrud, M. T., Rascati, K. L., Wilson, J. P., & Jensen, P. S. (2005). Trends in the use of typical and atypical antipsychotics in children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(6), 548–556.

Pelham, W. E., Foster, E. M., & Robb, J. A. (2007). The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. *Journal of Pediatric Psychology, 32*, 711–727.

Resko, S. (2011). Values talk exacerbates discrimination. *Hastings Center Report, 41*(2), S12.
Sadler, J. Z. (2007). The politics of psychiatry. *Project Syndicate*. http://www.project-syndicate.org/commentary/sadler2/English. Accessed 20 Jan 2014.
Sadler, J. Z., Jotterand, S. C. L. F., & Inrig, S. (2009). Can medicalization be good? Situating medicalization within bioethics. *Theoretical Medicine and Bioethics, 30*, 411–425.
Safer, D. J., Zito, J. M., & Fine, E. M. (1996). Increased methylphenidate usage for attention deficit disorder in the 1990s. *Pediatrics, 98*(6 Pt 1), 1084–1088.
Schlander, M. (2008). The NICE ADHD health technology assessment: A review and critique. *Child and Adolescent Psychiatry and Mental Health, 2*(1), 1.
Singh, I. (2004). Doing their jobs: Mothering with Ritalin in a culture of mother-blame. *Social Science and Medicine, 59*, 1193–1205.
Teich, J. L., & Buck, J. A. (2007). Mental health benefits in employer-sponsored health plans, 1997–2003. *The Journal of Behavioral Health Services & Research, 34*, 343–348.
Timimi, S., & Taylor, E. (2004). ADHD is best understood as a cultural context: For and against. *British Journal of Psychiatry, 184*, 8–9.
U.K. National Institute for Health and Clinical Excellence (NICE). (2008). *Attention deficit hyperactivity disorder: Diagnosis and management of ADHD in children, young people and adults*. http://www.nice.org.uk/guidance/index.jsp?action=byID&r=true&o=11632. Accessed 20 Jan 2014.
Wakefield, J. C. (2013). The DSM-5 debate over the bereavement exclusion: Psychiatric diagnosis and the future of empirically supported treatment. *Clinical Psychology Review, 33*, 825–845.
Whittington, C. J., Kendall, T., Fonagy, P., Cottrell, D., Cotgrove, A., & Boddington, E. (2004). Selective serotonin reuptake inhibitors in childhood depression: Systematic review of published versus unpublished data. *The Lancet, 363*(9418), 1341–1345.
Zito, J. M., Safer, D. J., dosReis, S., Gardner, J. F., Magder, L., Soeken, K., Boles, M., Lynch, F., & Riddle, M. A. (2003). Psychotropic practice patterns for youth: A 10-year perspective. *Archives of Pediatrics and Adolescent Medicine, 157*, 17–25.
Zito, J. M., Derivan, A. T., Kratochvil, C. J., Safer, D. J., Fegert, J. M., & Greenhill, L. L. (2008). Child and adolescent psychiatry and mental health. *Child and Adolescent Psychiatry and Mental Health, 2*, 24.

Prediction of Antisocial Behavior

107

Andrea L. Glenn, Farah Focquaert, and Adrian Raine

Contents

Abstract

A growing body of evidence suggests that biological factors such as genes, hormone levels, and brain functioning are associated with antisocial behavior. As research progresses, we will likely develop a better understanding of how biological factors very early in life influence the development of antisocial traits. This attempt to identify genes and early biological indicators of

A.L. Glenn (✉)
Center for Prevention of Youth Behavior Problems, Department of Psychology, The University of Alabama, Tuscaloosa, AL, USA
e-mail: Andrea.L.Glenn@ua.edu

F. Focquaert
Bioethics Institute Ghent, Department of Philosophy and Moral Sciences, Ghent University, Ghent, Belgium
e-mail: Farah.Focquaert@ugent.be

A. Raine
Departments of Criminology, Psychiatry, and Psychology, Jerry Lee Center of Criminology, University of Pennsylvania, Philadelphia, PA, USA
e-mail: araine@sas.upenn.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_149,

a propensity for criminal behavior raises a number of ethical concerns. In this chapter, the idea of an early destiny to crime is refuted and an attempt is made to explain the limited role that biological research alone has in predicting future criminal acts while the important potential that it has for helping us solve this significant societal problem is emphasized. It is argued that current knowledge on biological risk factors does not allow us to predict with reasonable certainty whether an individual is going to commit a crime. However, biological information may provide useful information about which individuals may be at somewhat greater risk for antisocial behavior, and thus may provide for the opportunity to intervene with programs designed to reduce this risk. A discussion of the ethical implications of this research is made at the end, including the potential for false positives, concerns about stigma, and the role of informed consent and the rights of children and parents.

Introduction

Some have suggested that biological research on crime may open the door to discrimination based on genes or brain indices, or that it may lead to individuals being labeled or punished before they have committed any crime. Media headlines such as "Criminal Behavior May Be Hard-Wired in the Brain, Researchers Find" (*Los Angeles Times*, November 17, 2009) perpetuate the idea that the identification of biological influences on criminal behavior indicates that some individuals are irreversibly destined for a life of crime. It is argued that many of the strong reactions to biologically based research may stem from a misunderstanding of the purpose of this type of research and the limitations of it in crime prediction. It is important to note that studies examining biological risk factors, like studies of environmental risk factors, represent *average* differences between groups. This means that there is no one-to-one relationship between biological factors and crime; many individuals with risk factors will not go on to commit crime, whereas some individuals who do not demonstrate a particular risk factor may proceed to commit crime. In other words, none of the studies produce large enough effects that alone would accurately predict who will grow up to be an offender. Furthermore, in most cases – as with social constructs – it has not been established that these biological risk factors are causal. Finally, the concept of multifinality comes into play. A wide range of biological and environmental factors may contribute risk for antisocial behavior. Thus, not all antisocial individuals will demonstrate the same biological (or environmental) deficits.

A start is made with an overview of genetic, electrophysiological, neuroendocrine, and neuroimaging measures that have been identified as early predictors of antisocial behavior. What the research in these domains can and cannot tell us about the risk for criminal behavior will be discussed.

Genetics

The idea that antisocial or criminal behavior is heritable is one that has been very controversial over the years. It has been alleged that examining the genetic factors that may contribute to crime is similar to "the kind of racist behavior we saw on the part of Nazi Germany" (Palca 1992). This reaction reflects a misunderstanding about the relationship between genes and behavior, and also a failure to recognize the potential for genetic research in aiding the development of positive outcomes, such as improved prevention and intervention methods.

Twin and adoption studies have provided significant scientific evidence that there is a substantial genetic contribution to criminal behavior (Moffitt 2005; Raine 2008). Genes are thought to account for approximately 50 % of the variation in criminal behavior across individuals in the population (Ferguson 2010). However, there is no single gene, or even a small group of genes that currently enables us to predict which individuals will commit crime in the future. The association between a single gene and a behavioral tendency or trait is typically very small (Canli and Lesch 2007). This is due to the fact that the pathway from a single gene to behavior is complex, and countless additional factors are introduced at each step. For example, environmental factors can alter the way that genes are expressed (e.g., turning genes "on" or "off"). This suggests that being a carrier of a particular gene does not necessarily mean that the functions of that gene will be realized. For example, animal studies have found that separating rat pups from their mother during the first 3 weeks of life results in increased expression of a gene associated with stress hormones in the hippocampus and prefrontal cortex, two brain regions which are critically involved in regulation of the stress response. Behaviorally, the rats demonstrated fearlessness and a blunted stress response in adulthood (Weaver et al. 2006), characteristics which are observed in some antisocial individuals and are thought to contribute to reduced sensitivity to punishment and blunted emotions. This documents that environmental factors early in life, such as maternal behavior, can directly change gene expression, thus altering the way in which the brain develops, and altering subsequent behaviors. Thus, the link between any particular gene and a behavior depends on the environmental context. Furthermore, the functioning of a particular gene may also depend on the presence of other genes.

The idea that the contribution of a single gene to complex behavioral traits is very small was demonstrated in a recent genome-wide association (GWA) study (Viding et al. 2010). GWA studies scan the genome for polymorphisms that may be more common in one group (e.g., individuals with persistent antisocial behavior) versus another (e.g., individuals without antisocial tendencies). Viding et al. (2010) used this method to attempt to identify genes that may be more common in children scoring high on both callous-unemotional traits and antisocial behavior, both of which are predictive of later aggression and delinquency (Frick et al. 2003). The study, which was powered to detect genes of large effect size, did

not find any genetic associations that reached genome-wide statistical significance (i.e., accounting for more than 1 % of the variance). This suggests that there is not going to be a single gene that will account for a large proportion of the variance in criminal behavior.

In addition to studies looking across the whole genome, studies have also examined the effects of specific gene variants that are hypothesized to be associated with antisocial behavior. These studies require very large sample sizes in order to detect any effect, because the effect of a single gene accounts for such a small proportion of the variance in the behavior (Lesch et al. 1996). In addition, many of the gene variants examined are very common in the population. For example, 20–35 % of individuals may be carriers of a particular "risk" genotype. Although the gene may significantly increase the risk for something such as persistent antisocial behavior, the majority of individuals with this particular gene variant do not become persistent criminals. Furthermore, many of these findings are limited to particular circumstances (e.g., the finding only applies to males, or only to individuals of low socioeconomic status). For example, Caspi et al. (2002) found that a variant of the monoamine oxidase A (MAOA) gene increased risk for antisocial behavior only in individuals who had experienced severe maltreatment as children. This gene-environment interaction emphasizes the idea that many factors are involved in the pathway from genes to behavior. Some genes may essentially sensitize children to environmental insults. Because the context in which genes operate varies from person to person, it would be not possible to use genetic information alone to predict whether an individual will commit a crime. However, what we may be able to do is identify youth who may be particularly sensitive to environmental risk factors, and try to change their environment in positive ways.

In sum, the genome overall exerts a substantial influence on criminal behavior, but no single gene is a major determinant. Whether any particular gene contributes to the risk for criminal behavior depends on the influence of environmental factors and other genes. Environmental factors such as poor nutrition, severe parental neglect, sustained physical and sexual abuse, early head injuries, learning disabilities, a family history of mental illness, poor cognitive functioning, toxin exposure, and a lack of parental supervision may all interact with genetic predispositions to increase the risk for antisocial behavior. By enriching the environment and/or reducing the presence of negative environmental factors such as childhood abuse, we may be able to reduce the influence of these genes on the development of traits related to antisocial behavior.

Psychophysiology and Neuroendocrinology

A number of studies have found that biological indicators early in life are predictive of later criminal behavior. For example, in a longitudinal study, Gao et al. (2010) compared individuals who had a criminal conviction at age 23 with individuals who did not have a conviction. Looking back at the results obtained 20 years earlier,

these two groups differed at age 3 on a test of fear conditioning, that is, the degree to which they developed a fear response to neutral stimuli that were repeatedly paired with aversive stimuli. Those with a conviction demonstrated relatively poorer fear conditioning at age 3 than those who did not. This suggests that poor fear conditioning at age 3 may act as a risk factor for future criminal behavior. Additional predictors of aggressive behavior were also observed in this sample. For example, low heart rate at age 3 years was found to predict aggression at age 11 years in children from high but not low social classes (Raine et al. 1997). Individuals with higher skin conductance orienting responses at age 3 were found to score higher on a self-report psychopathy scale at age 28 (Glenn et al. 2007). Psychopathic traits consistently have been found to predict offending (Grann et al. 1999; Gretton et al. 2004; Salekin et al. 2003).

Biological factors in adolescence have also been found to be predictive of later antisocial behavior. Tarter et al. (2009) found that increased levels of the hormone testosterone at ages 10–12 predicted assaultive behavior at ages 12–14, norm-violating behavior at age 16, and cannabis use at age 19. In another longitudinal study on delinquent male adolescents, attenuated heart rate response and stronger heart rate variability response to stress predicted higher reoffending rates 5 years later (de Vries-Bouw et al. 2011). Using electroencephalogram (EEG), Raine et al. (1990) found that indicators of early attentional processing at age 15 predicted criminality status at age 24. Arseneault et al. (2000) showed that minor physical anomalies, indicators of fetal neural maldevelopment, measured at age 14 years in 170 males predicted violent, but not nonviolent, delinquency at age 17 years.

These studies demonstrate that many individuals who go on to commit crime demonstrate biological differences early in life. By understanding how these biological risk factors interact with social and environmental risk, we may be better able to understand the development of criminal behavior.

Brain Imaging

Because brain imaging is a relatively new method, to the authors' knowledge, no studies to date have used the prospective longitudinal design discussed in the previous section to examine the neurobiological differences early in life that may confer risk for later criminal behavior. Thus far, the majority of studies conducted have been cross-sectional, examining differences in the brain in antisocial individuals compared to controls. As research currently stands, in the absence of gross impairments due to head injury or disease, claims about individual subjects cannot be made. Although effect sizes in brain imaging studies may be quite large, cut-off points at which it would be said that a particular individual demonstrates "abnormality" in a particular brain region are not yet determined. However, as research progresses, it may be able to quantitatively identify more specific neurobiological differences that increase the risk for antisocial behavior. By building up normative databases, it may be possible to create objective measures of brain structure or functioning that indicate a particular level of risk for antisocial behavior.

Like environmental and other biological risk factors, although it is unlikely that brain imaging could be used to predict *specific* criminal events, it is plausible that it could provide information about risk of recidivism. Indeed, two recent studies have found that information gained from brain imaging was predictive of future offending. In a sample of adult offenders, Aharoni et al. (2013) found that low activity in the anterior cingulate during performance on an inhibitory task was associated with increased odds of the individual being rearrested within a 4-year period. In a community sample, Pardini et al. (2013) found that men with lower amygdala volume at age 26 exhibited higher levels of aggression and violence at a 3-year follow-up. These initial studies demonstrating that brain structure and function may add valuable information in the prediction of future offending raise a number of ethical questions, which will be discussed in the next section.

One problem with using brain imaging in crime prediction, however, is that there is no strong evidence that the brain abnormalities that are observed at one time point will persist in the future. Some evidence suggests that there is stability in activity in the prefrontal cortex in prisoners over a 4-year period (Anckarsater et al. 2007). However, there is also considerable evidence that the environment can change the brain. Research suggests that the environment plays a significant role in shaping how the brain develops and how it functions, even in adulthood, meaning that there is potential for change. While this makes crime prediction more difficult, an advantage is that we have the ability to make changes for the better. For example, studies have shown that cognitive and behaviorally based treatments can produce lasting changes in brain functioning in patients with other types of mental health problems. Felmingham et al. (2007) found that after 8 weekly sessions of exposure-based therapy, brain functioning was changed in the anterior cingulate and amygdala in patients with posttraumatic stress disorder (PTSD). Similarly, Paquette et al. (2003) found that abnormal activity in the dorsolateral prefrontal cortex and parahippocampal gyrus was reduced to the level of controls in a sample of participants with spider phobia. In contrast to the idea that antisocial behavior may be "hard-wired" in the brain, these studies suggest that we can make potentially long-lasting changes in brain functioning through various forms of treatment. Future research examining the stability of structural and functional brain imaging findings in individuals will be helpful in determining how much weight to give this type of evidence, particularly in cases in which the imaging is conducted well after the crime has been committed.

Brain imaging may also be particularly useful in identifying subgroups of antisocial individuals who may be more or less responsive to particular forms of treatment. Research on other disorders has found that brain imaging is useful in understanding why some individuals respond better to treatment. For example, approximately half of patients with PTSD do not respond to cognitive behavioral therapy, which is the treatment of choice for the disorder. Bryant et al. (2008) found that pretreatment levels of functioning in the amygdala and anterior cingulate, as measured by fMRI, predicted which patients would respond to treatment. Greater amygdala and anterior cingulate activity was associated with poorer improvement after treatment. Assessing levels of brain functioning similarly may help us to

identify individuals who are at risk for antisocial behavior, and provide information about whether a particular form of intervention may be helpful to that individual. As research progresses, we may be able to identify youth who are particularly susceptible to antisocial behavior, and provide intervention programs that are more effective at targeting the specific neurobiological deficits of the individual.

Ethical Issues

Suppose that, in the future, we are able to identify with 80 % certainty that a child or young adult will develop persistent, violent behavior, either by genetic tests, brain imaging, or other biological tests. Suppose further that we have treatment programs available that may reduce the risk by 30 % or more, and hence decrease the chances that innocent individuals will be victimized. Under such circumstances, would it ethically be permissible to mandate screening and intervention in early life? Who should decide whether or not an at-risk child should receive treatment? Also controversial is the issue of prenatal genetic screening, embryo selection, and reproductive liberty. According to Ronald Dworkin (1993), we have a right to procreative autonomy, which he defines as "a right [of individuals] to control their own role in procreation unless the state has compelling reason for denying them that control" (p. 148). Does our right to procreative autonomy include the right "to reproduce with the genes we choose and to which we have legitimate access" (Harris 1998)? Should parents be allowed to have a child that has a high genetic risk for violent, antisocial behavior? Should they have sole decision-making authority whether or not to have the child?

John Harris (1998) asks us to imagine the following: We have six pre-implantation embryos awaiting transfer, and genetic screening reveals that three of these embryos have genetic diseases and three are normal. Which three embryos should be implanted? Do we have moral reasons to prefer those without the genetic diseases? Let us change the scenario a bit and imagine that we have six pre-implantation embryos awaiting transfer and genetic screening reveals that three of the embryos have a genetic predisposition to violent, antisocial behavior. All else being equal, do we have moral reasons to prefer the implantation of the embryos without the genetic predisposition? What if there was only one viable pre-implantation embryo and genetic screening revealed a predisposition to violent, antisocial behavior? Imagine that we have the technology to safely alter a genetic predisposition to violence by prenatally manipulating the genetic variant(s) that contribute to this predisposition. Do we have moral reasons to genetically manipulate the embryo in question? Although these normative questions are obviously not on the agenda at present, they might become more relevant in the future if scientific progress in the field of genomic preventive medicine continues to grow.

At present, early screening and intervention may benefit some individuals at risk of developing a particular disease, but carries a substantial risk of increasing anxiety and stigma, and falling prey to unnecessary interventions. As mentioned, not all individuals who exhibit an increased biological risk for a specific disorder

will develop that disorder. Furthermore, the environment interacts with biological (including genetic) factors, and therefore should be fully considered. Pressing ethical dilemmas that we will be faced with when thinking about early screening for risk factors related to the development of mental disorders (e.g., genetic and biological predictors for violent, antisocial behavior) are related to: (1) the efficacy, cost-effectiveness, and likely "false alarm" rate of population-based screening strategies, (2) anxiety and stigma, (3) informed consent and the obligations between parents and children.

1. From a public health ethics perspective, population-wide screening of individuals (genetic or otherwise) is only justifiable if an efficacious and cost-effective intervention exists to prevent the development of the disorder in individuals who are identified as at-risk. However, even if efficacious interventions are available, we do not know whether screening and early intervention will be more cost-effective than treating those who actually have the disease or display disease symptoms. Moreover, it is unclear whether information about genetic risk is sufficient to motivate behavior changes and whether this information will motivate individuals enough to seek and continue treatment. It is possible that information about genetic or biological risk could undermine an individual's confidence in his/her ability to change his/her behavior for the better, and may therefore contribute to deviant behavior rather than preventing it. Alternatively, such knowledge combined with effective treatments to ameliorate biological risk factors could motivate antisocial individuals to seek treatment.

   In general, public health professionals have raised worries that population-wide genomic screening may draw our attention away from *more* effective population-based prevention strategies (e.g., prenatal and/or school-based nutrition programs, programs against child abuse and maltreatment, programs aimed at socioeconomic risk factors, etc.) that aim at lowering the risk for the entire population and not merely for a normatively defined at-risk group. When do we consider an individual to be at-risk, or to have a high versus low risk? The answer to this question involves normative judgments concerning health versus ill-health. Population-wide predictive screening and intervention strategies will involve false negatives, leaving out individuals in need of treatment, and false positives, leading to unnecessary interventions that may be considered harmful to the individual in question (Holm 2007). Interventions based upon population-based screening should not be a substitute for the implementation of policies that reduce exposure to common risk factors in the entire population.

2. Stigma reduces an individual to someone who is "tainted" and should be "avoided." Stigma can have many negative consequences by limiting access to material resources and opportunities and lowering psychological well-being. Through a process of social exclusion, stigma often aggravates diseases by discouraging individuals to seek treatment and by worsening treatment outcomes for those who are in treatment. Stigma can have long-term detrimental effects on the self-esteem and self-worth of stigmatized individuals and has been described as an assault on human dignity (Bayer 2008; Spriggs et al. 2008).

Should the harm that is caused by stigmatization be avoided at all cost? Do we have strong moral reasons to prioritize the potential harm that is caused by stigmatization over the potential benefits that may result from screening and effective interventions? Although stigmatization has the potential to significantly impact an individual's human dignity, it may be argued that the primacy of non-maleficence ("do no harm") is not absolute (Shickle and Chadwick 1994). According to Shickle and Chadwick (1994), there must be room for a trade-off between the harm that results from screening and the harm that is brought about by a failure to screen. However, while we may have legitimate reasons to opt for screening (preventing harm), we may also have legitimate reasons ("do no harm," e.g., stigma) to refrain from screening. Although we must acknowledge responsibility for the harm that is caused by a failure to screen (and thus failure to prevent), we should not implement every screening program that could conceivably prevent harm. Peter Singer (1979) proposes the following test: "If we can prevent something bad without sacrificing anything of comparable moral significance, we ought to do it" (169).

Spriggs et al. (2008) argue that a greater justification for screening children is needed for disorders that carry a greater risk of stigmatization. The higher the genetic determination of a particular disorder and the fewer the modifiable genetic determinants, the higher is the risk of stigmatization. In those cases, stigmatization is linked to mistaken neuro-essentialist beliefs that reduce an individual to his/her genetic or biological makeup and equate genetic or biological risk factors with "un-changeability" or "un-treatability." At the same time, Spriggs et al. (2008) predict that screening and intervention for mental disorders in early life is most likely to be of use for mental disorders with a high genetic determination. Indeed, screening and early intervention is unlikely to be of great value in disorders with a low genetic determination due to the unreliability of future disease risk. Based upon a cost-benefit analysis of genetic testing and an overall assessment of the reasons for and against genetic testing, Spriggs et al. (2008) argue that early screening for genetic (or biological) risk for mental disorders is not warranted unless there is a family history of mental disease, and unless efficacious, cost-effective, low-risk interventions exist to preventively reduce the risk. The authors do not make a difference between disorders that may involve violent, aggressive behavior toward others, such as antisocial personality disorder, and disorders that mainly involve self-harm. If effective treatment programs are available, could screening for violent, antisocial predictive factors in early life be warranted for children without a (known) family history? Do the potential benefits in terms of public safety outweigh the potential risks in terms of stigmatization and unnecessary interventions?

3. Imagine that a child, without ever having committed a violent crime, has been identified as at-risk for violent, antisocial behavior based upon genetic or biological screening and imagine further that treatment programs exist that may reduce this risk substantially. Who should decide if the child enters the treatment program? Should the parents have sole decision-making authority or should the child be involved in the decision-making process? "Child

protectionism" is the view that children, although they may have rights, are cognitively and emotionally incapable of deciding for themselves and, hence, of exercising their rights. Therefore, either their parents or legal guardians, in combination with professionals or legal authorities, should decide for them based on their current and future best interests. Child protectionists argue that too much weight has been placed on children's autonomy, and that this might be detrimental for children rather than empowering (Ross 1998). In contrast, "child liberationists" argue that children's rights should be respected and that children's views should be given due weight in accordance with their age and maturity. Lynn Hagger (2009) argues that parents have the duty to gradually diminish their proxy decision-making and allow their maturing children to make their own choices. By providing opportunities to choose for themselves, children can develop and learn the necessary skills to become competent adults. A shared decision-making model may be preferred because (1) this creates the kind of environment in which the child's developing autonomy is respected and the child's decision-making skills are nurtured, and because (2) shared decision-making has been found to result in better treatment outcomes and coping behavior (Focquaert 2011). As is the case with most interventions targeting behavior change, successful treatment will depend to a large degree on the willingness of the individual in question to undergo treatment. It is therefore of the utmost importance that the child is involved in the decision-making process and, preferentially, that a *shared* decision-making process can be established where the parent or legal guardian gives his/her informed consent and the child his/her assent.

Conclusions and Future Directions

Children have the physical, cognitive, and emotional means of being physically aggressive toward others by 1 year of age. Individual differences in the frequency and severity of this aggression can be observed shortly thereafter (Tremblay 2008). Unsurprisingly, there are biological differences that can be observed in children at very young ages, and some of these differences have been linked to increased risk for antisocial behavior. However, it is unlikely that genetic and neurobiological information will ever be precise enough to allow us to predict with very high probability that an individual will commit a specific act. Thus, concerns that biological information may eventually be used to predict crime before it occurs are premature.

A more likely scenario is that biological information may be used to identify which individuals are at somewhat greater *risk* for traits associated with criminal behavior. As discussed, there are a number of ethical considerations related to early screening and intervention that arise, including the substantial risk of increasing anxiety and stigma, and the issue of whether interventions should be mandatory. It is argued that a shared decision-making process involving both the parent and the child may produce the best outcomes.

One of the most difficult neuroethical challenges is finding the sensitive balance between protecting society and protecting our individual civil rights. With these issues in mind, it is suggested that biological research on criminal behavior should be viewed as part of the solution rather than part of the problem. Understanding the biological factors that influence criminal behavior puts us one step closer to understanding how this behavior develops, and will help us to create more targeted and sophisticated methods for prevention and intervention, which is the ultimate goal. These interventions often have benefits that extend beyond the goal of preventing antisocial behavior, and include improvements in school performance, social adjustment, and reduced risk of substance abuse and other mental health problems. By sensibly and cautiously integrating genetic and neuroscience findings with public policy, future crime and violence can be better prevented.

Cross-References

▶ Determinism and Its Relevance to the Free-Will Question
▶ Ethical Issues in the Neuroprediction of Addiction Risk and Treatment Response
▶ Free Will, Agency, and the Cultural, Reflexive Brain
▶ Neuroimaging Neuroethics: Introduction
▶ Neuroscience, Free Will, and Responsibility: The Current State of Play

References

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., Gazzaniga, M. S., & Kiehl, K. A. (2013). Neuroprediction of future rearrest. *Proceedings of the National Academy of Sciences, 110*(15), 6223–6228. doi:10.1073/pnas.1219302110.

Anckarsater, H., Piechnik, S., Tullberg, M., Ziegelitz, D., Sorman, M., Bjellvi, J., . . . Forsman, A. (2007). Persistent regional frontotemporal hypoactivity in violent offenders at follow-up. *Psychiatry Research, 156*(1), 87–90. doi:10.1016/j.pscychresns.2006.12.008.

Arseneault, L., Tremblay, R. E., Boulerice, B., Seguin, J. R., & Saucier, J. F. (2000). Minor physical anomalies and family adversity as risk factors for violent delinquency in adolescence. *American Journal of Psychiatry, 157*, 917–923.

Bayer, R. (2008). Stigma and the ethics of public health: Not can we but should we. *Social Science & Medicine, 67*(3), 463–472. doi:10.1016/j.socscimed.2008.03.017.

Bryant, R. A., Felmingham, K., Kemp, A., Das, P., Hughes, G., Peduto, A., & Williams, L. (2008). Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder. *Psychological Medicine, 38*(4), 555–561. doi:10.1017/s0033291707002231.

Canli, T., & Lesch, K.-P. (2007). Long story short: The serotonin transporter in emotion regulation and social cognition. *Nature Neuroscience, 10*, 1103–1109.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., . . . Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. *Science, 297*, 851–854.

de Vries-Bouw, M., Popma, A., Vermeiren, R., Doreleijers, T. A. H., Van de Ven, P. M., & Jansen, L. M. C. (2011). The predictive value of low heart rate and heart rate variability during stress for reoffending in delinquent male adolescents. *Psychophysiology, 48*, 1596–1603.

Dworkin, R. (1993). *Life's dominion: An argument about abortion, euthanasia, and individual freedom*. New York: Knopf.

Felmingham, K., Kemp, A., Williams, L., Das, P., Hughes, G., Peduto, A., & Bryant, R. (2007). Changes in anterior cingulate and amygdala after cognitive behavior therapy of posttraumatic stress disorder. *Psychological Science, 18*, 127–129.

Ferguson, C. J. (2010). Genetic contributions to antisocial personality and behavior: A meta-analytic review from an evolutionary perspective. *The Journal of Social Psychology, 150*(2), 160–180. doi:10.1080/00224540903366503.

Focquaert, F. (2011). Pediatric deep brain stimulation: Parental authority versus shared decision-making. *Neuroethics*. doi:10.1007/s12152-12011-19098-12154.

Frick, P. J., Cornell, A. H., Barry, C. T., Bodin, S. D., & Dane, H. E. (2003). Callous-unemotional traits and conduct problems in the prediction of conduct problem severity, aggression, and self-report of delinquency. *Journal of Abnormal Child Psychology, 31*(4), 457–470.

Gao, Y., Raine, A., Venables, P. H., & Dawson, M. E. (2010). Association of poor childhood fear conditioning and adult crime. *American Journal of Psychiatry, 167*, 56–60.

Glenn, A. L., Raine, A., Venables, P. H., & Mednick, S. (2007). Early temperamental and psychophysiological precursors of adult psychopathic personality. *Journal of Abnormal Psychology, 116*(3), 508–518.

Grann, M., Långström, N., Tengström, A., & Kullgren, G. (1999). Psychopathy (PCL-R) predicts violent recidivism among criminal offenders with personality disorders in Sweden. *Law and Human Behavior, 23*(2), 205–217. doi:10.1023/A:1022372902241.

Gretton, H. M., Hare, R. D., & Catchpole, R. E. H. (2004). Psychopathy and offending from adolescence to adulthood: A 10-year follow-up. *Journal of Consulting and Clinical Psychology, 72*(4), 636–645. doi:10.1037/0022-006X.72.4.636.

Hagger, L. (2009). *The child as vulnerable patient: Protection and empowerment*. Farnham, UK: Ashgate.

Harris, J. (1998). Rights and reproductive choice. In J. Harris & H. Soren (Eds.), *The future of human reproduction: Choice and regulation* (pp. 5–37). Oxford: Oxford University Press.

Holm, S. (2007). Obesity interventions and ethics. *Obesity Reviews, 8*, 207–210. doi:10.1111/j.1467-789X.2007.00343.x.

Lesch, K.-P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B., Petri, S., . . . Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science, 274*, 1527–1531.

Moffitt, T. E. (2005). The new look of behavioral genetics in developmental psychopathology: Gene-environment interplay in antisocial behaviors. *Psychological Bulletin, 131*, 533–554.

Palca, J. (1992). NIH wrestles with furor over conference. *Science, 257*, 739.

Paquette, V., Levesque, J., Mensour, B., Leroux, J. M., Beaudoin, G., Bourgouin, P., & Beauregard, M. (2003). "Change the mind and you change the brain": Effects of cognitive-behavioral therapy on the neural correlates of spider phobia. *NeuroImage, 18*, 401–409.

Pardini, D. A., Raine, A., Erickson, K., & Loeber, R. (2013). Lower amygdala volume in men is associated with childhood aggression, early psychopathic traits, and future violence. *Biological Psychiatry*. doi:10.1016/j.biopsych.2013.04.003.

Raine, A. (2008). From genes to brain to antisocial behavior. *Current Directions in Psychological Science, 17*, 323–328.

Raine, A., Venables, P. H., & Williams, M. (1990). Relationships between N1, P300, and contingent negative variation recorded at age 15 and criminal behavior at age 24. *Psychophysiology, 27*(5), 567–574.

Raine, A., Venables, P. H., & Mednick, S. A. (1997). Low resting heart rate age 3 years predisposes to aggression at age 11 years: Evidence from the Mauritius Child Health Project. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 1457–1464.

Ross, L. F. (1998). *Children, families and health care decision making*. Oxford: Clarendon.

Salekin, R. T., Ziegler, T. A., Larrea, M. A., Anthony, V. L., & Bennett, A. D. (2003). Predicting dangerousness with two Millon Adolescent Clinical Inventory psychopathy scales: The

importance of egocentric and callous traits. *Journal of Personality Assessment, 80*(2), 154–163. doi:10.1207/S15327752JPA8002_04.

Shickle, D., & Chadwick, R. (1994). The ethics of screening: Is 'screeningitis' an incurable disease? *Journal of Medical Ethics, 20*(1), 12–18.

Singer, P. (1979). *Practical ethics*. Cambridge: Cambridge University Press.

Spriggs, M., Olsson, C. A., & Hall, W. (2008). How will information about the genetic risk of mental disorders impact on stigma? *The Australian and New Zealand Journal of Psychiatry, 42*(3), 214–220. doi:10.1080/00048670701827226.

Tarter, R. E., Kirisci, L., Gavaler, J. S., Reynolds, M., Kirillova, G., Clark, D. B., . . . Vanyukov, M. (2009). Prospective study of the association between abandoned dwellings and testosterone level on the development of behaviors leading to cannabis use disorder in boys. *Biological Psychiatry, 65*(2), 116–121. doi:10.1016/j.biopsych.2008.08.032.

Tremblay, R. E. (2008). Understanding development and prevention of chronic physical aggression: Towards experimental epigenetics studies. *Philosophical Transactions of the Royal Society B, 363*, 2613–2622.

Viding, E., Hanscombe, K. B., Curtis, C. J. C., Davis, O. S. P., Meaburn, E. L., & Plomin, R. (2010). In search of genes associated with risk for psychopathic tendencies in children: A two-stage genome-wide association study of pooled DNA. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 51*, 780–788.

Weaver, I. C. G., Meaney, M. J., & Szyf, M. (2006). Maternal care effects on the hippocampal transcriptome and anxiety-mediated behaviors in the offspring that are reversible in adulthood. *Proceedings of the National Academy of Sciences, 103*, 3480–3485.

Mind, Brain, and Education: A Discussion of Practical, Conceptual, and Ethical Issues

108

Daniel Ansari

Contents

Abstract

Recent years have seen a tremendous growth in efforts to connect rapidly growing insights into how the brain learns to the field of education. Different names have been given to such efforts, including "Educational Neuroscience" and "Mind, Brain, and Education." The aim of this chapter is to discuss these recent efforts and to provide an overview of the conceptual, practical, and ethical challenges faced by these novel, transdisciplinary efforts. To do so, the chapter begins with an overview of some examples of recent research efforts to connect research on Mind, Brain, and Education (MBE). To do so, the chapter begins with an overview of some examples of recent research in the emerging field of Mind, Brain and Education. Specifically, the chapter reviews evidence from the study of the neurocognitive processes of typical and atypical reading development in an effort to illustrate the merit of MBE. This is followed by a discussion of the conceptual and practical challenges that MBE needs to consider, such as the level at which evidence from the study of neurocognitive processes can influence education and how neuroscientists and educators can play complementary roles in the

D. Ansari
Numerical Cognition Laboratory, Department of Psychology & Brain and Mind Institute, The University of Western Ontario, London, ON, Canada
e-mail: daniel.ansari@gmail.com; daniel.ansari@uwo.ca

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_146,

construction of the field. Against this background, the chapter considers ethical challenges, including the need to effectively and accurately communicate evidence, to carefully consider the commercialization of neuroscience, including the use of brain stimulation, to enhance cognitive functions and for classroom application.

Introduction

Recent decades of neuroscience research have witnessed an unprecedented growth in our understanding of the neural mechanisms that subserve human learning. Significant advances in our understanding of the neuronal machinery that allows humans to acquire complex skills such as reading, writing, mathematics, and problem solving have been made. Furthermore, our understanding of brain development and how experience changes brain function and structure has grown tremendously (Johnson 2001; Lenroot and Giedd 2006; Munakata et al. 2004). We are learning not only how the brain comes to process the content of the most classical domains of education, such as numeracy and literacy, but insights into the brain mechanisms underlying motivation, attention, and working memory are constraining our understanding and modelling of how children acquire new skills that are essential to their success in society.

Against the background of these advances in basic research, there have been growing calls to apply knowledge from neuroscience to education. On the surface, it is a “no-brainer” that neuroscience should be capable of informing education. Without our brains, we cannot learn and therefore understanding the machinery (the brain) that underpins the ability of children and adults to learn in the context of educational settings should have implications for how to improve educational environments and pedagogy. A key concept here is “neuronal plasticity” – the ability of the brain to change as a function of experience. Early neurophysiological studies revealed that sensory deprivation or enrichment changes the brains of animals, showing that experience shapes the brain (Buonomano and Merzenich 1998). Today, we can study the effects of complex environmental and experiential differences, such as cross-cultural variability, on brain structure and function (Ansari 2012). Mounting research suggests that the brain is more plastic than we originally thought (though within constraints) and that our brains continue to be capable of functional and structural change (plasticity) in adulthood, which has potential implications for life-long learning, a topic of much interest in many aging Western societies (May 2011).

Plasticity is key to education. Teachers are the orchestrators of their students’ neuronal plasticity during classroom time. In order for knowledge to be acquired, the brain has to encode information, which involves changes in the connectivity between nerve cells (i.e., synaptic plasticity). The brain is not a static organ, but instead dynamically adapts to the environment. Education is a process of inducing brain plasticity through instruction in a social context. When children learn, for example, to read (for a greater discussion of this topic, see below), their brain changes from them seeing letters on the page as meaningless characters to the point at which they read with a flashlight at night and use these previously meaningless characters to follow stories and create meanings in their minds.

Notwithstanding strong critics (Bruer 1997), over the past 15 years or so, the enthusiasm for a new science of learning and education that combines insights from cognitive science, neuroscience, and psychology has grown exponentially. Terms such as "Neuroeducation," "Educational Neuroscience," and "Mind, Brain, and Education" have been used to describe these efforts. The author of the present chapter prefers the term "Mind, Brain, and Education" (henceforth MBE) because it reflects the mutual and interactive influence of research from cognitive psychology, neuroscience, and educational research. New organizations such as the International Mind, Brain and Education Society (IMBES; www.imbes.org) and journals (e.g., "Trends in Neuroscience and Education" and "Mind, Brain and Education") have been launched to attract research and theory that lies at the intersections between neuroscience, education, cognitive science, and psychology." Interest in such new directions has extended beyond the realm of academia. Organizations such as the Organization for Economic Cooperation and Development (OECD) have convened expert groups to examine the relationship between Mind, Brain, and Education in an effort to find ways to improve education in OECD membership countries across the globe (OECD 2002). In addition, new graduate programs are being launched in many departments, the first of which was the Master's in Mind, Brain and Education at the Harvard Graduate School of Education founded by Dr. Kurt Fischer (http://www.gse.harvard.edu/academics/masters/mbe/).

As can be seen from the above, the field of MBE is growing rapidly. As with any new field, there are many conceptual, practical, and ethical issues to consider in order to facilitate the successful growth of the field. The aim of this chapter is to provide a critical reflection on and discussion of the potential of Mind, Brain, and Education. The chapter will commence with a discussion of examples from neuroscience research that are having implications for education and to illustrate some of the findings from neuroscientific research that are fueling the enthusiasm for Mind, Brain, and Education. The review of these findings will also consider how evidence from cognitive neuroscience might impact education now and in the future. The chapter will then turn to conceptual and practical challenges that are faced by this new, translational field. Against the background of having considered what MBE research looks like and what challenges face the field, the chapter will discuss the ethical challenges that the emerging field of Mind, Brain, and Education is and will be confronted with.

Examples of Success in "Mind, Brain, and Education"

The aim of the following section is to provide the reader with some examples of recent interdisciplinary research that demonstrates the potential of the emerging field of Mind, Brain, and Education. This review does not aim to be comprehensive, but simply to illustrate through a few examples the potential significance of research that bridges the gap between cognitive neuroscience and education.

Perhaps, the best example of how neuroscience research is shaping education comes from research in the domain of reading and the breakdown of reading abilities (i.e., Developmental Dyslexia). Noninvasive brain imaging methods such as functional Magnetic Resonance Imaging (fMRI) and Event-Related Potentials (ERPs) have enabled researchers to map out the brain regions involved in reading and to image the time course of neural activity associated with the process of reading. Through the use of such methods, it has become possible not only to study which brain regions are involved during reading, but how these change over the course of learning and development as well as what differences exist between the brain structure of typical readers and those with Developmental Dyslexia (for reviews, see: Gabrieli (2009); Schlaggar and McCandliss (2007)). The aim here is not to provide an exhaustive review of these studies but to highlight some findings that show how neuroscience can be used to constrain educational problems and inform educational practice. One area in which neuroscience is making great strides is in the prediction of reading success and failure. For example, several studies using ERPs to record the brain responses of neonates and infants, have revealed that infants brain responses to sounds predict individual differences in reading years later (Guttorm et al. 2001, 2010; Molfese 2000; Pihko et al. 1999). These are powerful data for a number of reasons. First of all, they reveal that the pre-reading brain of infants who will go on to experience difficulties in reading respond differently to those who will develop normal literacy skills. This draws attention to the early scaffolds of reading, resulting in many potential implications for early diagnosis and remediation. Secondly, these kinds of data are difficult to obtain with any other measure traditionally used in behavioral research with young infants and children, thereby demonstrating the added value of using neuroimaging methods.

In studies with older children using both structural and functional neuroimaging, it has been demonstrated that structural variables, such as brain volume and white matter integrity, as well as functional measures of brain activation during reading-related tasks (e.g., rhyming), predict significant variability in children's reading scores. One might argue that this is hardly surprising and that it is far more cost-effective to use traditional behavioral measures as predictor variables. However, Hoeft et al. demonstrated that the combined use of behavioral and neuroimaging measures as predictors of reading (specifically decoding skills) explains significantly more variance than either measure used in isolation (Hoeft et al. 2007). Thus, neuroimaging and behavioral measures each explain unique variance, which leads to overall better prediction of reading outcomes. These data speak against the notion that neuroimaging does nothing in addition to what can be gleaned from behavioral measures alone, but instead shows that such measures explain outcomes that cannot be captured by behavior alone.

In a more recent study, Hoeft et al. showed another powerful way in which neuroimaging measures can be used to predict educationally meaningful outcomes. The authors asked whether neuroimaging data could predict who will go on to show gains in reading performance. In addition to a large battery of behavioral tests of reading, writing, and IQ, structural and functional brain imaging data was acquired from children with and without Developmental

Dyslexia. The same children were tested again 2.5 years later. Behavioral data suggested that while some children with Developmental Dyslexia exhibited significant gains in reading abilities, another group of children demonstrated no change on behavioral tests of reading competencies. By using the behavioral and neuroimaging data acquired at the outset of the study to predict who ended up showing reading gains compared to children who did not, the authors were able to show that structural and functional neuroimaging measures were able to predict which children ended up exhibiting gains in their reading abilities. In striking contrast, none of the behavioral variables were able to statistically predict which children exhibited gains (Hoeft et al. 2011). These findings provide strong evidence for the possibility of "neuroprognosis" and also demonstrate that, in some cases, neuroimaging measures (both structural and functional) may be a more sensitive way in which to predict later outcomes (given that the behavioral data did not exhibit such predictive power). In addition to providing strong evidence to suggest that neuroimaging measures can be used to predict change in educationally relevant outcomes (e.g., gains in reading ability), Hoeft et al. found that individual differences in the structure and function of the right inferior frontal cortex were particularly predictive of such gains. This is an interesting finding, as reading is mostly associated with a left-lateralized network of brain activity and structure. However, there have been other studies suggesting that the right hemisphere plays an important role in response to intervention and may represent compensatory neural mechanisms (Temple et al. 2003). Thus, individuals who are better able to use these right-lateralized compensatory mechanisms may show greater gains in reading ability. This finding provides a significant constraint on our understanding of the mechanisms that drive individual differences in improvements in reading abilities. It is not as though we necessarily see the normalization of disrupted brain circuits, but instead it appears that the recruitment of regions not typically associated with reading is what is associated with the recovery of impaired reading skills. This is not only important from the point of view of understanding the mechanisms underlying the recovery of reading abilities, but it may also help to constrain how reading interventions are designed. In other words, future efforts may be directed at better understanding what mechanisms drive the recruitment of such compensatory neural processes and how these might be optimally harnessed. In this way, neuroimaging provides novel constraint on recovery of cognitive function and by revealing the mechanisms can guide future intervention programs.

Importantly, this is not an isolated case of "neuroprognosis." Very recently, Supekar et al. (2013) showed a similar result for the prediction of response to intervention in the domain of math learning. These researchers demonstrated that measures of hippocampal volumes as well as functional connectivity between the hippocampus and other brain regions were a significant predictor of individual differences in response to a structured intervention for children with math learning difficulties. Furthermore, convergent with the data from reading reviewed above, the behavioral data acquired prior to the intervention were not able to predict individual differences in the response to intervention. Thus, these data

provide convergent evidence for the potential power of neuroprognostics from two domains: reading and mathematics.

Another recent neuroscientific study of reading further illustrates the value added of neuroscience in addressing educational questions. One of the most hotly debated topic in special education is the use of so-called discrepancy criteria (Stanovic, 2005). Many researchers and practitioners define children as having specific difficulties in domains such as reading, writing, and arithmetic if they have below the normal range scores on the domain of interest but perform within the normal range on other tests of abilities. So, for example, according to the discrepancy criteria approach, a child with Developmental Dyslexia can only be defined as such if their reading abilities are both well below their general academic abilities and if their non-reading abilities are within the normal range. There have been many arguments against such stringent discrepancy criteria with researchers suggesting that children who have reading difficulties that are either discrepant or non-discrepant from their other intellectual and academic abilities have the same educational needs and are indeed indistinguishable from one another when it comes to measures of their reading ability (Fletcher et al. 1992). In a recent study, neuroimaging was used to constrain the question of whether children with and without a discrepancy between their reading and IQ scores showed different patterns of brain activation (Tanaka et al. 2001). The authors compared the brain activation of children with and without a reading-IQ discrepancy while they performed a reading-related task during functional neuroimaging (fMRI). Both univariate and multivariate analyses of the functional imaging data clearly demonstrated that both groups of children exhibited indistinguishable patterns of under activation (relative to non-impaired controls) of areas in the left hemisphere that are typically associated with successful reading. These findings show that the neurobiology underpinning reading does not differ between children with and without a reading-IQ discrepancy and therefore place a significant novel constraint on the special education debate surrounding the utility of using discrepancy based criteria to identify children with specific learning difficulties.

Beyond the domain of reading, there are many other examples in other educationally relevant areas that can be characterized as success stories in the emerging field of "Mind, Brain, and Education." For example, several studies, ranging from experiments with rats to functional neuroimaging studies with humans, have demonstrated the powerful effect that socioeconomic status and early experiences have on brain structure and function (Hackman et al. 2010). These data have revealed how early experiences and stress embed themselves into our biology and have long-term consequences for our health, well-being, and ability to learning. Without going into the details of the individual studies that demonstrate these powerful effects of early experience and socioeconomic status, these findings clearly have very important implications for education. They demonstrate the importance of early educational programs, especially in disadvantaged communities, in mitigating the profoundly negative effects of early, adverse experiences.

Conceptual and Practical Challenges

The above section provides a non-exhaustive discussion of some select examples (primarily from the study of the neurobiology of typical and atypical reading acquisition) of successful research in the emerging field of “Mind, Brain, and Education.” While such advances illustrate that there lies much promise in such transdisciplinary research, there are also many conceptual and practical challenges that lie in the path of such efforts. The aim of the following section is to discuss some of these challenges and potential ways of navigating such obstacles.

Whenever different fields of both inquiry and application meet one another, a need for a common language and set of expectations emerges. One expectation that is prevalent when neuroscience and education meet is that neuroscience will have a direct impact on classroom instruction (Ansari and Coch 2006). In other words, scientists generate evidence, communicate this evidence to educators who will then go on to apply it. This expectation will inevitably lead to disappointment, as results from research cannot be directly applied, but instead need to be gradually translated through an iterative process that needs to involve multiple bidirectional interactions between the research laboratory and the classroom and requires individuals who are well versed in both cognitive neuroscience and education to facilitate the process of translation (Ansari and Coch 2006; Varma et al. 2008).

Translation of evidence does not involve the presentation of recipes by researchers to educators, but must involve a collaborative effort that allows for the establishment of a common language and sets of expectations. In other words, to ensure successful transdisciplinary interactions between cognitive neuroscience and education, it is critical for expectations to be realistic and that a broader conceptualization of potential translations of insights from research into practice be adopted. Take the examples of studies, discussed above, revealing brain signals measured in neonates during language processing predict literacy impairments many years later. One may ask some of the following questions upon being presented with this evidence: what to do with this evidence in terms of practice? Does it mean that we can intervene very early in development or does it mean that the fate of children is determined from an early age onward and cannot be changed? If we can intervene, then how and when? What is the diagnostic utility of the event-related potential measurements for individual babies? Is it reliable enough to serve as a diagnostic tool? When considering these questions, it quite rapidly becomes apparent that research generates questions whose answers may lead to translation, but that the raw evidence often cannot be directly translated into practical applications. Evidence, whether from neuroscience or other scientific disciplines, is the beginning of a translational process.

The expectation that research will be directly translated into application also reflects another common assumption: insights from the research laboratory will invariably *change* education. Visions of entirely new classrooms come to mind that are “brain friendly” or a complete revolution in the way in which certain subjects are currently taught. While it is indeed possible that new evidence will lead to

changes in education, evidence can also play a critical role in affirming as well as speaking against what educators already do today and thereby strengthen certain approaches they adopt in their pedagogy every day. In this way, evidence may help to provide empirical grounding to certain educational approaches and techniques. In the above discussion of research on the neuroscience of dyslexia, powerful examples of how neuroscience can inform educational decision and help to arbitrate between different approaches was discussed.

The study by Tanaka et al. demonstrates that the neural correlates of reading impairments do not differ between dyslexics who were diagnosed using a reading-IQ discrepancy criteria and those who were not (who therefore showed impairments in reading and other cognitive functions). This confirms that discrepancy criteria used to diagnose children with developmental dyslexia do not appear to isolate a group of children who have qualitatively different neuronal deficits in reading from those who also have non-reading difficulties (Tanaka et al. 2011). This study therefore can be used to inform educational decision-making, in this case the method of diagnosing children with reading impairments. It does not change the way in which diagnoses are made, but instead provides critical information to guide decision-making. The use of evidence to confirm or disconfirm certain pedagogical choices is a powerful way by which evidence from the cognitive neuroscience laboratory can influence education and may indeed initially be the most realistic way in which education and cognitive neuroscience can be connected in the context of educational practice (Thomas 2013).

A similar role of neuroscientific evidence was recently highlighted by Laurence Steinberg in a discussion of the role that neuroscience has played in US supreme court decisions regarding the criminal culpability of adolescents (Steinberg 2013; see also Johnson & Giedd, this volume). What Steinberg argues is that neuroscience most likely had an influence on the legal decision-making not by revealing something completely new that no other previous evidence, such as behavioral evidence, had shown, but by confirming common sense, intuition, and evidence from behavioral evidence. In other words, neuroscientific evidence played an important supportive role in the process of legal decision-making.

One might argue that if the role of scientific evidence, such as data from cognitive neuroscience, plays only a supportive role, then one might as well ignore it. This argument, however, ignores the fact that confirmation of one set of opinions or intuitions is often accompanied by the rejection of alternative approaches where there either does not exist evidence to support their efficacy or evidence that demonstrate that such approaches are inefficacious. In this way, evidence provides a means by which to inform education. Rather than basing educational decisions on opinions and intuitions, a culture of evidence-based education uses evidence to make informed decisions.

It follows from the above discussion that, if teacher education programs were to systematically train teachers in the evaluation of empirical knowledge during preservice training, then such individuals would be better equipped to use evidence to inform their pedagogical decision-making. This would allow teachers to become

critical consumers of empirical evidence from a range of field of inquiry relevant to education, including but not restricted to evidence from cognitive psychology and neuroscience, and to seek out evidence to both confirm and reject particular educational approaches (Ansari 2005). Training teachers and other educational professionals in the language of science, how evidence is generated and evaluated, may hold the key to effective processes of translation. In the same way, cognitive neuroscientists generating evidence, which they hope will inform education, should become versed in pedagogy and approaches in educational practice and research. This will lay the foundation for translation that is supported by individuals from a multitude of backgrounds to ensue.

Ethical Challenges

Having considered some, but certainly not all, of the practical and conceptual challenges faced by the field of Mind, Brain, and Education, the present discussion now turns to the ethical challenges that this field of inquiry has encountered as well as those that might lie in its future paths. This discussion will not consider ethical challenges associated with the use of neuroimaging methods to study the developing brain (for an excellent discussion of such challenges, see: Coch 2007). Instead the present section will consider broad ethical challenges that lie at the conceptual, rather than the methodological, level of Mind, Brain, and Education as a new field of inquiry.

Recent years have seen an impressive increase in the public's awareness and interest in neuroscience. Popular magazines are full of articles about the brain, and neuroscience is enjoying an unprecedented amount of public attention. While such attention is broadly welcome (at least from the vantage point of neuroscientists), it does also pose some significant challenges that require the attention of neuroscientists. One of these is the creating of a knowledge hierarchy. Specifically, because of the great attention paid to neuroscience, its informational merit, relative to other sources of evidence, may be judged too highly. As discussed above, the role of neuroscience is most frequently a complementary one. In other words, neuroscientific evidence can inform other areas, such as education, in conjunction with insights from other disciplines, such as the behavioral sciences. A set of recent studies have demonstrated that the great value assigned to neuroscientific evidence may not always lead to good conclusions about that evidence. One of these examined the influence of brain images on how individuals evaluate scientific evidence. Noninvasive imaging methods, such as functional Magnetic Resonance Imaging (fMRI), generate aesthetically pleasing images of brain "activation." These images, of course, are not a direct representation of "activity" but rather represent color-coded statistical maps showing the probability of the "activation" (which in turn is represented by changes in blood flow that are thought to correlated with neuronal activity) in certain brain regions. These images provide powerful illustrations of data, but it turns out that their presence can also have a significantly negative influence on the way in which non-experts evaluate neuroscientific evidence (Weisberg et al. 2008). Specifically, Weisberg et al. presented both naïve

adults, students in a neuroscience course and neuroscience experts with explanations of psychological phenomena. These explanations were varied in two ways: (1) They were either good explanations or invalid explanations. (2) They were either paired with or without neuroscientific evidence; however, this evidence was completely irrelevant to the logic of both good and bad explanations. The investigators found that this irrelevant neuroscientific evidence influenced the way in which the group of naïve participants and the neuroscience students judged the explanations. Specifically, these participants were more satisfied with explanations that were paired with irrelevant neuroscientific evidence and, perhaps most alarmingly, they were less likely to recognize the bad explanations (i.e., found them more satisfying than bad explanations without irrelevant neuroscientific data contained within the explanation) when these were paired with irrelevant neuroscientific evidence. In other words, the presence of neuroscientific evidence appeared to make non-experts think more highly of bad explanations of scientific phenomena.

The study by Weisberg et al. is a powerful explanation of how the current "hype" around neuroscience could lead to biases in the way in which individuals evaluate the validity of scientific evidence. However, it is very important to consider that the study by Weisberg and colleagues and related investigations of the effect of brain images on scientific evaluations (McCabe and Castel 2008) have recently not been replicated (Michael et al. 2013) in large-scale studies and therefore, it is unclear how reliable these effects are and how specific they are to brain images, as opposed to other stimuli that could bias the judgment of scientific data (Farah and Hook 2013).

Regardless of whether the results are reliable and are specific to neuroscience, there are some general ethical implications for researchers in the field of MBE that go beyond the particulars of this research. Specifically, researchers in this emerging field of inquiry need to recognize that their understanding of the strengths and limitations of neuroscientific evidence is not necessarily at the same level of individuals who are not trained in neuroscience and therefore the way in which evidence is communicated becomes a critical ethical issue for researchers in MBE, particularly those who are generating neuroscientific evidence. Researchers in MBE must be aware of how their research results may be perceived by non-experts, such as teachers, and must therefore ensure that their findings are communicated in ways to avoid misunderstandings and the inappropriate use of the evidence. Researchers must think carefully about how they talk not only about the evidence they have generated but also about its implications for the classroom and to avoid overstating the potential for direct application. For example, teachers may not be aware that neuroscientific data is often based on averages and thus, a given finding may not be representative of every student or that must studies in cognitive neuroscience are conducted with adults participants and that the data may not be readily generalizable to children. Therefore, neuroscientists need training in how to communicate to and engage with the public in ways that are ethical and thereby avoid the creation of misconceptions (Illes et al. 2010).

Effective communication and engagement is especially important when it comes to sharing neuroscientific evidence with educators. Many teachers are extremely enthusiastic about neuroscience and seek out opportunities to learn more about

neuroscience (Hook and Farah 2012). This enthusiasm represents a great opportunity for the field of Mind, Brain, and Education to grow and for dialogue between neuroscience and education to ensue. Without educators and "buy-in" from educators, the transdisciplinary field of MBE cannot succeed. However, such dialogue must be effective and avoid scenarios that lead teachers to construct misconceptions about neuroscience, or so-called neuromyths. Some so-called facts (neuromyths) about the brain, such as the idea that we only use 10 % of our brains or that some individuals are "left-brained" while others are "right-brained," appear to be very persistent among teachers (OECD 2002).

In a recent study of primary and secondary teachers in the Netherlands, it was found that around half of them believed neuromyths (Dekker et al. 2012). Moreover, and perhaps more alarming, the authors found that those teachers who were enthusiastic about neuroscience and read more popular science magazines were also those who were more likely to believe neuromyths. These data demonstrate the importance of providing teachers with better tools to evaluate scientific evidence in order to avoid misconceptions about neuroscientific (and other empirical) evidence that could be informative in terms of their pedagogy (Dubinsky 2010). Furthermore, it is important that researchers think carefully about how this evidence is going to be presented to teachers. The finding that those who read more popular science magazines were also more likely to believe in neuromyths provides a stark illustration that information itself is clearly not enough, but that the quality of that information and how it is delivered matters. It is likely that such training might be most efficient at the preservice level so that teachers enter their profession not only seeking evidence-based approaches to their pedagogy but also being capable of carefully evaluating evidence.

The ethical challenge, for both neuroscientists and teacher education institutions, of addressing the prevalence and proliferation of neuromyths among educators, is further exacerbated by the commercialization of so-called brain-based programs. These programs are often advertised as being based on neuroscientific data, though a closer examination often reveals that such data are, at best, tangentially related to the commercial programs that are being advertised. Furthermore, the majority of such programs have not been evaluated in adequately controlled trials and therefore, it is unknown to what extent they will be efficacious.

In a large-scale evaluation of computerized programs that purport to train certain brain functions, it was recently found that while training on such tasks improved task performance, there was no evidence of transfer of the trained abilities to other tasks, even if they involved very closely related neurocognitive functions (Owen et al. 2010). Thus, the promised effects of the "brain training" programs investigated did not hold. Furthermore, in some cases, the empirical evaluation of such programs reveals that they can actually lead to decrements in certain neurocognitive functions rather than improvements. In a study evaluating the effect of babies watching special "educational" programs on the infants language development, Zimmerman and colleagues found that greater consumption of such videos was associated with decrements rather than improvements in language development, as measured by a parental questionnaire regarding the child's communicative abilities (Zimmerman et al. 2007). This is, of course, not to say that all so-called brain

training programs will be either inefficacious or harmful, but these two studies demonstrate the importance of grounding such applications in empirical research and making sure that educators demand such evidence before investing financial and pedagogical resources into their application.

Educators who are not trained in evaluating products and asking questions about the evidence base behind them, but who are, at the same time, highly enthusiastic about neuroscience may buy into such programs without reservations. To avoid the proliferation of such “brain-based” approaches in classrooms, it is important that individuals, such as administrators and classroom teachers who have the responsibility of choosing such programs, are trained in how to judge the efficacy of an educational program as well as the evidence base (or lack thereof) of any program that is being offered to them. The same applies to books written on “brain-based” education. If neuroscientists and teacher educators do not take action to prevent the spread of empirically unsupported programs and books, then this could lead to the further proliferation of neuromyths and the application of untested programs that may have no effect at all or, even worse, detrimental effects on neurocognitive functions. The field of MBE now needs an infrastructure for such training and ethical information sharing to take place. Training teachers and educational decision makers is going to be insufficient in preventing the proliferation of “brain-based” programs that have no supporting evidence. More and more such programs will flood the market as interest in neuroscience and its application to education increases. Neuroscientists and educators should therefore work together with politicians, funding agencies, and regulators to come up with guidelines. Eventually, given the critical role that education plays in society, there should be regulatory organizations, similar to those that exist to regulate medical products.

Beyond considering neuroscientific evidence that has implications for education by virtue of showing how learning in certain domains affects brain structure and function, there exists a growing body of research that explores attempts to stimulate the brain in order to enhance cognitive functions (e.g., Cohen Kadosh et al. 2010; Hauser et al. 2013). Such studies are suggesting that the application of weak electrical current to the brain via scalp electrodes can lead to enhancements of cognitive functions through the induction of plastic changes in the neuronal architectures supporting these functions. These studies have received widespread attention in the media and it is foreseeable that commercialization of such approaches will take place shortly, if it is not already underway. One could imagine then, that it will not be too long before schools are offered packaged brain stimulators for their students that are to be applied during instruction to improve learning. This therefore raises a whole host of ethical considerations related to the potential, undesired, and uncontrollable side effects of such stimulation.

In this context, it is important to consider that while the data suggesting that methods such as Transcranial Direct Current Stimulation (TDCS) are intriguing, the existing studies are exclusively conducted with comparatively small samples of adult participants, therefore making their generalizability to samples of children impossible to estimate. Children’s brains are developing and cannot be readily compared with those of adults. Therefore, the response of children’s brains to

stimulation may be quite different from what is currently being observed in adults. Moreover, though many study show improvements on behavioral measures following brain stimulation during learning, there also exists evidence to suggest that such stimulation can also have negative effects on cognitive processes (Iuculano and Cohen Kadosh 2013). Thus, it is clear that brain stimulation does not only affect the neurocognitive functions it is aiming to target but leads to side effects, the extent and magnitude of which is unknown. Generally, brain stimulation studies are very much in their infancy and little is known about the mechanisms through which brain stimulation exerts its effects on learning. Given this, researchers have an ethical responsibility to communicate about the potential of such methods to be applied in the classroom with caution. It is likely that such approaches and similar ones that try to modulate neurocognitive functions that are relevant to educational processes via stimulation or pharmacology are going to rapidly increase over the coming years and thus this will be a major frontier of neuroethics for MBE.

On a broader level, it is the contention of this author that MBE must keep in mind that education is a deeply cultural activity. As such, education is not a fixed entity. Education varies across historical time, contexts, and cultures. The priorities of what children should be learning changes with the general shifts in our societies and cultures. It is brain plasticity that allows for these rapid adaptations to changing sociocultural demands and contexts.

Often when discussing the role of neuroscience in education, the metaphor of a muscle is used. Specifically it is argued that the brain is like a muscle that needs to be exercised and that therefore education is a form of exercising your brain. In the opinion of the present author, this metaphor is overly simplistic and its implications may be misleading. While there is a range of ways in which you can exercise the muscles in your body, it does not by any stretch of the imagination resemble the diversity of how and what humans learn. Learning and education are not simply the acts of exercising a muscle in a particular way that can be replicated across cultures and contexts.

Though neuroscience has great potential to provide information that could improve education and help us decide which educational approaches are most optimal (through systematic, evidence-based evaluation), it will never be able to inform us as to what we should be teaching our children. Neuroscience is and should be agnostic as to the specific goals of education. This will remain the purview and responsibility of societies. Therefore, the discussion around issues in MBE and the neuroethics of it should consider that efforts to bring neuroscience to bear on educational problems do not lead to normalization of education across time and context and a fixed view of what education means and what children should and should not learn.

Summary and Conclusions

Evidence from neuroscience is increasingly discussed in the context of education. How can we use our growing insights into the mechanisms by which the brain learns across domains to improve education? This is a hotly debated question today.

On the one hand, there is much to be enthusiastic about. The study of education is in many ways the study of brain plasticity. If our brains were unable to change in response to experience, education would not be possible. Every month, new studies are published that show how the brain changes as a function of learning across different domains, such as reading and mathematics. On the other hand, while the evidence base is steadily growing, efforts to translate such evidence into educational practice face significant conceptual, practical, and ethical challenges.

There is a need to move beyond models of translation that focus on the direct translation of neuroscience laboratory research to the classroom and instead engage in a more iterative process of translation that involves neuroscientists, educational researchers, and practitioners. Furthermore, beyond changing education, evidence from neuroscience *in conjunction* with evidence from other domains has an important role to play in supporting what is already being implemented in education (thereby lending evidence to approaches that have previously not been evidence-based) and in using evidence to inform educational decision-making. The role of empowering educational professionals and decision makers through giving them tools to be critical consumers of evidence is discussed as both a conceptual and ethical challenge. Neuroscientists need to assume greater responsibility in ensuring that the information they generate is interpreted correctly and must guard against the proliferation of neuromyths, premature commercialization, and the use of neuroscientific approaches to enhance neurocognitive functions that have not been systematically tested in pediatric populations and for which side effects are currently understudied.

Finally, while there is much reason to be enthusiastic about the emergence of MBE, it is important that society continues to discuss the direction of education independently of neuroscientific evidence. While evidence from neuroscience can inform which approaches work and why and perhaps help to optimize them, education and educational priorities remain deeply cultural activities.

Cross-References

► Ethical Implications of Brain Stimulation
► Human Brain Research and Ethics
► Neuroenhancement
► Neuroimaging Neuroethics: Introduction
► Research in Neuroenhancement

References

Ansari, D. (2005). Time to use neuroscience findings in teacher training. *Nature, 437*(7055), 26.

Ansari, D. (2012). Culture and education: New frontiers in brain plasticity. *Trends in Cognitive Sciences, 16*, 93–95. doi:10.1016/j.tics.2011.11.016.

Ansari, D., & Coch, D. (2006). Bridges over troubled waters: Education and cognitive neuroscience. *Trends in Cognitive Sciences, 10*(4), 146–151. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16530462.

Bruer, J. T. (1997). Education and the brain: A bridge too far. *Educational Researcher, 26*, 4–16.
Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience, 21*, 149–186. doi:10.1146/annurev.neuro.21.1.149.
Coch, D. (2007). Neuroimaging research with children: Ethical issues and case scenarios. *Journal of Moral Education, 36*, 1–18. doi:10.1080/03057240601185430.
Cohen Kadosh, R., Soskic, S., Iuculano, T., Kanai, R., & Walsh, V. (2010). Modulating neuronal activity produces specific and long-lasting changes in numerical competence. *Current Biology: CB, 20*(22), 2016–2020. doi:10.1016/j.cub.2010.10.007.
Dekker, S., Lee, N. C., Howard-Jones, P., & Jolles, J. (2012). Neuromyths in education: Prevalence and predictors of misconceptions among teachers. *Frontiers in Psychology, 3*, 429. doi:10.3389/fpsyg.2012.00429.
Dubinsky, J. M. (2010). Neuroscience education for prekindergarten-12 teachers. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(24), 8057–8060. doi:10.1523/JNEUROSCI.2322-10.2010.
Farah, M. J., & Hook, C. J. (2013). The Seductive Allure of "Seductive Allure". *Perspectives on Psychological Science, 8*(1), 88–90. doi:10.1177/1745691612469035.
Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, S. E., & Shaywitz, B. A. (1992). The validity of discrepancy-based definitions of reading disabilities. *Journal of Learning Disabilities, 25*(9), 555–561. 573. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1431539.
Gabrieli, J. D. (2009). Dyslexia: A new synergy between education and cognitive neuroscience. *Science, 325*(5938), 280–283. doi:325/5938/280 [pii] 10.1126/science.1171999.
Guttorm, T. K., Leppänen, P. H., Richardson, U., & Lyytinen, H. (2001). Event-related potentials and consonant differentiation in newborns with familial risk for dyslexia. *Journal of Learning Disabilities, 34*, 534–544.
Guttorm, T. K., Leppänen, P. H. T., Hämäläinen, J. A., Eklund, K. M., & Lyytinen, H. J. (2010). Newborn event-related potentials predict poorer pre-reading skills in children at risk for dyslexia. *Journal of Learning Disabilities, 43*, 391–401.
Hackman, D. A., Farah, M. J., & Meaney, M. J. (2010). Socioeconomic status and the brain: Mechanistic insights from human and animal research. *Nature Reviews Neuroscience, 11*, 651–659.
Hauser, T. U., Rotzer, S., Grabner, R. H., Mérillat, S., & Jäncke, L. (2013). Enhancing performance in numerical magnitude processing and mental arithmetic using transcranial Direct Current Stimulation (tDCS). *Frontiers in Human Neuroscience, 7*, 244. doi:10.3389/fnhum.2013.00244.
Hoeft, F., Ueno, T., Reiss, A. L., Meyler, A., Whitfield-Gabrieli, S., Glover, G. H., . . . Gabrieli, J. D. (2007). Prediction of children's reading skills using behavioral, functional, and structural neuroimaging measures. *Behavioral Neuroscience*, *121*(3), 602–613. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17592952
Hoeft, F., McCandliss, B. D., Black, J. M., Gantman, A., Zakerani, N., Hulme, C., . . . Gabrieli, J. D. E. (2011). Neural systems predicting long-term outcome in dyslexia. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 361–366. doi:10.1073/pnas.1008950108.
Hook, C. J., & Farah, M. J. (2012). Neuroscience for educators: What are they seeking, and what are they finding? *Neuroethics, 6*(2), 331–341. doi:10.1007/s12152-012-9159-3.
Illes, J., Moser, M. A., McCormick, J. B., Racine, E., Blakeslee, S., Caplan, A., . . . Weiss, S. (2010). Neurotalk: Improving the communication of neuroscience research. *Nature Reviews Neuroscience*, *11*(1), 61–69. doi:10.1038/nrn2773.
Iuculano, T., & Cohen Kadosh, R. (2013). The mental cost of cognitive enhancement. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(10), 4482–4486. doi:10.1523/JNEUROSCI.4927-12.2013.
Johnson, M. H. (2001). Functional brain development in humans. *Nature Review Neuroscience, 2*(7), 475–483. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11433372.

Lenroot, R. K., & Giedd, J. N. (2006). Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioural Reviews, 30*(6), 718–729. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16887188.

May, A. (2011). Experience-dependent structural plasticity in the adult human brain. *Trends in Cognitive Science, 15*, 475–482.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*(1), 343–352. doi:10.1016/j.cognition.2007.07.017.

Michael, R. B., Newman, E. J., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non)persuasive power of a brain image. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-013-0391-6.

Molfese, D. L. (2000). Predicting dyslexia at 8 years of age using neonatal brain responses. *Brain and Language, 72*, 238–245.

Munakata, Y., Casey, B. J., & Diamond, A. (2004). Developmental cognitive neuroscience: Progress and potential. *Trends in Cognitive Sciences, 8*(3), 122–128. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15301752.

OECD. (2002). *Understanding the brain: Towards a new learning science*. Paris: OECD Publishing.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., . . . Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*, 775–778.

Pihko, E., Leppänen, P. H., Eklund, K. M., Cheour, M., Guttorm, T. K., & Lyytinen, H. (1999). Cortical responses of infants with and without a genetic risk for dyslexia: I. Age effects. *Neuroreport, 10*, 969–973.

Schlaggar, B. L., & McCandliss, B. D. (2007). Development of neural systems for reading. *Annual Review of Neuroscience, 30*, 475–503. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17600524.

Stanovich, K. (2005). The future of a mistake: Will discrepancy measurement continue to make the learning disabilities field a pseudoscience? *Learning Disability Quarterly, 28* 103–106. Retrieved from http://www.jstor.org/stable/10.2307/1593604

Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience, 14*(7), 513–518. doi:10.1038/nrn3509.

Supekar, K., Swigart, A. G., Tenison, C., Jolles, D. D., Rosenberg-Lee, M., Fuchs, L., & Menon, V. (2013). Neural predictors of individual differences in response to math tutoring in primary-grade school children. *Proceedings of the National Academy of Sciences of the United States of America, 110*(20), 8230–8235. doi:10.1073/pnas.1222154110.

Tanaka, S., Inui, T., Iwaki, S., Konishi, J., & Nakai, T. (2001). Neural substrates involved in imitating finger configurations: An fMRI study. *Neuroreport, 12*(6), 1171–1174. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11338186.

Tanaka, H., Black, J. M., Hulme, C., Stanley, L. M., Kesler, S. R., Whitfield-Gabrieli, S., . . . Hoeft, F. (2011). The brain basis of the phonological deficit in dyslexia is independent of IQ. *Psychological Science*, *22*, 1442–1451. doi:10.1177/0956797611419521.

Temple, E., Deutsch, G. K., Poldrack, R. A., Miller, S. L., Tallal, P., Merzenich, M. M., & Gabrieli, J. D. (2003). Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America, 100*(5), 2860–2865. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12604786.

Thomas, M. S. C. (2013). Educational neuroscience in the near and far future: Predictions from the analogy with the history of medicine. *Trends in Neuroscience and Education, 2*(1), 23–26. doi:10.1016/j.tine.2012.12.001.

Varma, S., McCandliss, B., & Scwartz, D. (2008). Scientific and pragmatic challenges for bridging education and neuroscience. *Educational Researcher, 37*(3), 140.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477. doi:10.1162/jocn.2008.20040.

Zimmerman, F. J., Christakis, D. A., & Meltzoff, A. N. (2007). Associations between media viewing and language development in children under age 2 years. *The Journal of Pediatrics, 151*(4), 364–368. doi:10.1016/j.jpeds.2007.04.071.

Normal Brain Development and Child/Adolescent Policy

109

Sara B. Johnson and Jay N. Giedd

Contents

Abstract

In the last 25 years, magnetic resonance imaging technology has fundamentally changed how human brain development is conceptualized. Brain structures and the communication among them are now understood to change well into early adulthood in ways that impact maturity of judgment. The popular conversation

S.B. Johnson (✉)
Johns Hopkins University School of Medicine, Baltimore, MD, USA
e-mail: sjohnson@jhsph.edu

J.N. Giedd
National Institutes of Health, National Institute of Mental Health, Bethesda, MD, USA
e-mail: jg@nih.gov

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_148,

about where to draw the line between childhood and adulthood for policy purposes has highlighted a number of complex neuroethical issues including: balancing responsibility and autonomy, the strengths and frailties of human competence, and decision making in the era of neuroimaging. In this chapter, two public policy issues: Informed consent and legal culpability are used to illustrate the emerging neuroethical challenges and opportunities involved in using neuroscience to inform child and adolescent policy.

This chapter begins with an overview of historical attempts to use biological benchmarks of adult maturity. This historical perspective is followed by an introduction to the neuroethical issues involved in informed consent and legal culpability for adolescents, and the brain and behavioral science that has been brought to bear on these policy questions. The focus of this scientific review is the development and deployment of the cognitive capacities that are the foundation of maturity of judgment during late childhood and adolescence: self-control, inhibition, emotion regulation, and vulnerability to peer influence. Finally, the opportunities and potential pitfalls involved in using brain science to inform child and adolescent policy are considered.

Introduction

In the last 25 years, magnetic resonance imaging (MRI) technology has fundamentally changed the way the development of the human brain is conceptualized. It is now well established that brain structures and the communication among them change throughout childhood and adolescence, and indeed, across the life course (Giedd et al. 1999; Sowell et al. 1999, 2001). These new understandings have prompted re-examination of long-held ideas about what maturity means, both biologically and behaviorally. Particularly salient for health and legal policy is establishing the point in cognitive and socio-emotional development at which children become adults. Two areas of public policy have brought the developmental neuroscience of childhood and adolescence into sharp focus: informed consent and legal culpability. These policy issues have exposed inconsistencies among sociological, political, and biological definitions of maturity, and have shaped the public conversation about where to draw the line between childhood and adulthood.

This chapter begins with an overview of historical attempts to use biological benchmarks of adult maturity. This historical perspective is followed by an introduction to the neuroethical issues involved in informed consent and legal culpability for adolescents, and the brain and behavioral science that has been brought to bear on these policy questions. The focus of the scientific review is the development and deployment of self-control, inhibition, emotional regulation, and peer influence during late childhood and adolescence. Finally, the opportunities and potential pitfalls involved in using brain science to inform child and adolescent policy are considered.

Defining Maturity

The line between childhood and adulthood, while indistinct from a developmental point of view, is clear for the purposes of informed consent and legal culpability. Adulthood is the point at which individuals are considered competent both to make autonomous choices about their well-being, and they are held legally responsible for those choices. Throughout history, a number of biological benchmarks have been used to codify the skills needed to successfully participate in the tasks of adulthood. For example, in thirteenth century feudal England, the age of majority (i.e., legal recognition of adulthood) was raised from 15 to 21, citing the physical strength needed to bear the increasing heft of metal protective armor and the skill required to fight on horseback (James 1960). Absent such tangible physical requirements, contemporary legal definitions of adulthood are arguably developmentally arbitrary, and certainly less consistent. The age of consent for marriage has varied historically and currently varies widely by country (from puberty to age 21) and within the United States. For example, in New Hampshire, the marriageable age with parental consent is 13 for females and 14 for males, and 18 for both sexes without parental consent. In Nebraska, both males and females must be at least 17 with parental consent and 19 without parental consent. Age of consent for sex is now between 16 and 18 in all US states, but in 1880 ranged from ages 7 to 12, with most states setting the minimum age at 10 (Robertson 2008). Minimum ages for involvement in the political process are also nonuniform. Although the voting age in the USA has been 18 since 1972, the minimum age to hold office varies from 18 for mayor, 21 for governor, 25 for Governor or Representative, 30 for Senate, and 35 for President. Age of conscription is 18, drinking alcohol 21, tobacco 18, purchasing a rifle or shotgun 18, and a handgun 21. The range of these ages is at odds with a more fundamental biological notion of a distinct maturation level at which a person can make sound decisions on his or her behalf and participate independently in society.

Perhaps because they recognize the potential utility of a more developmentally informed approach, policy makers and judges have called on developmental scientists to weigh in on when adolescents can be expected to think and behave like adults. In this chapter, the policy issues of consent and culpability are considered to illustrate how new understandings of the “teen brain” have informed ethical questions about autonomy and justice at the threshold of adulthood.

Legal and Developmental Issues in Consent and Culpability

Informed Consent

Informed consent is a core principle of bioethics. The ethical precept of respect for persons dictates that cognitively competent adults should be free to make their own choices about their health and health care. In the USA, before age 18, consent to medical treatment generally requires parental consent; minors are assumed to be incompetent under the law. Instead, parents are charged with representing and

protecting the best interests of their children (Thomas and O'Kane 1998). The law does, however, recognize that adolescents stand, for the purposes of maturity, somewhere between adults and children (Ford et al. 2004). A "mature minor" is allowed to provide informed consent if he or she demonstrates "sufficient maturity to understand and appreciate the benefits and risks of the proposed medical treatment" (Derish and Heuvel 2000, p. 115). (It is worth noting that this is a *cognitive* definition of maturity). While current laws vary by jurisdiction, minors may consent to a subset of medical services without parental consent/notification. These services include: testing for sexually transmitted infections, contraception, prenatal care, drug and alcohol treatment, mental health services, and abortions (American College of Obstetricians and Gynecologists 2009). In addition to recognizing the adolescent's increasing autonomy, there is a public interest in removing barriers to contraception or treatment. Adolescents have among the highest rates of foregone health care, and concerns about privacy from parents have been shown to be a significant deterrent to seeking care (Ford et al. 1999, 2004).

In 1990, before longitudinal neuroimaging research demonstrated protracted development of the child and adolescent brain, the US Supreme Court grappled with the issue of adolescent maturity of judgment for the purposes of making autonomous health care decisions. In *Hodgson vs. Minnesota*, physician Hodgson contested the State of Minnesota's law that required both parents be notified before a minor could receive an abortion (United States Supreme Court 1990). Ultimately, the Court decided that only one parent must be notified, and established a judicial bypass process in which a judge could fill the role of the parent. The Court found that "the State has a legitimate interest in the welfare of its young citizens, whose immaturity, inexperience, and lack of judgment may sometimes impair their ability to exercise their rights wisely" (United States Supreme Court 1990). In a friend of the court brief, developmental scientists representing the American Psychological Association, citing substantial behavioral science research, argued that by the age of 14, adolescents have adult-like "abilities outlined in the law as necessary for understanding treatment alternatives, considering risks and benefits, and giving legally competent consent" (American Psychological Association 1989, p. 20). The developmental benchmark of maturity legally required for consent in this case was adult-like cognitive capacity in the context of medical decision-making (Steinberg et al. 2009).

More than a decade after *Hodgson*, another legal case brought issues of adolescent maturity back to the forefront, this time with evidence from developmental neuroscience brought to bear. Nevertheless, as the next section will make clear, a single, unifying, age-based definition of maturity is elusive, despite additional insight provided by the advent of neuroimaging.

Legal Culpability

The juvenile justice system in the USA was created at the turn of the twentieth century as a "therapeutic" alternative to the adult justice system. Premised on the idea that juveniles are immature, it was designed to shepherd individuals whose youthful

indiscretions led them astray safely into adulthood by addressing the underlying causes of their criminality (Feld 1993). Since the 1980s, however, there has been a shift to a more retributive, rather than rehabilitative approach to juvenile justice. Juveniles are increasingly charged as adults and punished accordingly (i.e., more harshly) (Feld 1993; Steinberg 2009). This shift has prompted intense scientific interest in how developmental science might inform culpability and judicial consequences for juveniles (Cauffman and Steinberg 2012). Regardless of age, in the criminal justice system, deficiencies in decision-making capacity or judgment (e.g., impulsivity, short-sightedness) and the presence of coercion each mitigate culpability and reduce the severity of the resulting penalty (Steinberg and Scott 2003). Thus, when considering culpability and consequences for adolescents, it is important to understand developmental changes in impulsivity, planning, foresight, and vulnerability to peer influence, particularly now that juveniles are more likely to find themselves in adult court, or a juvenile system that looks increasingly like adult court.

In 2005, the US Supreme Court Case, *Roper v. Simmons*, focused attention on the intersection of culpability and developmental science (United States Supreme Court 2005). Seventeen-year-old Christopher Simmons was convicted of murder during a robbery committed with friends and was sentenced to death. Simmons' legal team, bolstered by friend of the court briefs from developmental scientists, argued that 17-year olds' still-developing brains make them fundamentally different from adults in terms of legal culpability; consequently, the death penalty should not be imposed on those who commit crimes before age 18 (American Psychological Association and the Missouri Psychological Association 2004). The Court sided with Simmons and overturned the death penalty for juveniles, drawing parallels with prior case law that established those with intellectual disabilities as less culpable. The Court argued that juveniles' "lack of maturity and undeveloped sense of responsibility," coupled with greater vulnerability to and less control over their surroundings, reduced their blameworthiness. Further, while acknowledging the shortcomings of a categorical, age-based definition of maturity, the Court affirmed age eighteen as "the point where society draws the line for many purposes between childhood and adulthood" (United States Supreme Court 2005).

Since *Roper*, the US Supreme has reconsidered the issue of adolescent legal culpability in ways that acknowledge the changing landscape of developmental science. These cases reflect the increasing integration of structural and functional neuroimaging work into the narrative of adolescent immaturity of judgment. In 2010, the Supreme Court decided *Graham vs. Florida* (United States Supreme Court 2010). The case involved Terrance Graham, who was 16 when he committed armed burglary with peers and was given probation without judgment. He was subsequently arrested for additional crimes, his probation revoked, and sentenced to life in prison. On appeal, the Supreme Court overturned life without parole for non-homicide crimes. In its opinion, the Court concluded that "developments in psychology and brain science continue to show fundamental differences between juvenile and adult minds" and cited a friend of the court brief filed by the American Medical Association that highlighted the late maturation of "parts of the brain involved in behavior control" (United States Supreme Court 2010).

Adolescents' immaturity of judgment was again considered by the Supreme Court in *Miller vs. Alabama*, which examined mandatory life sentences for juveniles involved in homicide crimes (United States Supreme Court 2012). This opinion consolidated two state court cases, each of which involved 14-year-old boys who were involved in botched robberies that resulted in homicides. Kuntrell Jackson was convicted for his role in a burglary that resulted in another youth killing a video store clerk. Evan Miller killed his neighbor by setting his trailer on fire after he purchased drugs from him along with another youth. Each boy received a mandatory life sentence.

In its opinion in *Miller*, the Supreme Court required lower courts to consider "an offender's youth and attendant characteristics" in determining appropriate sentencing for juveniles convicted of homicide. The opinion acknowledged the role "science and social science" had previously played in the *Roper* and *Graham* decisions. Further, it cited accumulating scientific evidence in the years since *Roper and Graham* to support developmental risk-factors for criminality, particularly vulnerability to peer influence (Aber et al. 2012; United States Supreme Court 2012). Together, these three cases, *Roper, Graham*, and *Miller* highlight the increasing integration of behavioral and brain science into questions of adolescent culpability and responsibility.

Summary

While informed consent and capability are fundamentally different issues, they focus attention on the same question. When are adolescents' cognitive and socio-emotional capacities sufficiently mature to render them functionally equivalent to adults in terms of decision-making? In an effort to balance competing societal and legal demands, e.g., respecting the privacy of the family to make decisions free from government interference, preserving public safety, promoting public health, and holding individuals responsible for crimes they commit, policy makers and judges have increasingly turned to developmental neuroscience science for guidance. In *Hodgson*, the behavioral science was used to illustrate that adolescents are cognitively mature to make decisions about their health, while in *Roper, Miller*, *and Graham*, a combination of behavioral and neuroimaging science was used to demonstrate that teens are fundamentally different from adults when it comes to impulsivity, planfulness, sensitivity to peer influence, and the brain systems underlying those traits. This seeming inconsistency is explained by evidence that cognitive and socio-emotional processes reflect different domains of maturity with different developmental time courses. We turn now to the scientific evidence that helps to explain this phenomenon.

Brain, Behavior, and Maturity in Adolescence

While there is still much to learn from developmental neuroscience, one question is settled: When it comes the brain, adolescents are not just smaller adults – nor is the adolescent brain a broken or deficient adult brain. Although the adolescent brain has

different features than in childhood or adulthood, across species of social mammals, from rodents to primates, the adolescent brain is well suited to the evolutionary skills required for a successful transition to adulthood (Spear 2000). These include: risk tolerance and novelty seeking needed to venture away from the natal family and try new things, the social affiliation needed to create a new community with peers, and the drive to reproduce in order to pass along one's genes to the next generation (Spear 2000). Although these drives have individual and societal costs, viewed through an evolutionary lens, adolescents' brains are elegantly adaptive.

Normative Brain Changes from Childhood to Early Adulthood

Magnetic resonance imaging (MRI) combines a powerful magnet, radio waves, and sophisticated computer software to produce exquisitely accurate pictures of brain anatomy (i.e., anatomic MRI) and physiology (i.e., functional MRI) without the use of ionizing radiation. The safety of MRI allows not only the scanning of healthy children and adolescents but also repeated scanning throughout the course of development. These features, along with widespread availability of MRI technology, have led to thousands of studies that have dramatically deepened our understanding of the path and influences on brain development in health and illness. Brain tissue on anatomic MRI scans is usually categorized into two main types: (1) gray matter, which is comprised mostly of cell bodies, dendrites, and dendritic processes including synapses – the junctions between communicating brain cells (Braitenberg 2001); and (2) white matter, which is comprised mainly of axons wrapped in myelin – a fatty insulating substance that increases and regulates the speed of neuron-neuron communication.

A key insight provided by MRI studies is that the brain does not mature by becoming larger. It matures by becoming more specialized and more interconnected. By age 6, the brain is already approximately 93 % of its maximum size. The maximum size occurs around age 11 in girls and 13 in boys and then decreases slightly as the demands of the environment optimize brain circuitry – strengthening useful connections and eliminating unused or non-useful connections. The powerful twin processes of overproduction followed by selective elimination drive brain development from conception until death, but the net balance is what tips at around the time of puberty. The pruning process is important in at least two respects: It allows the brain to change in response to the specific demands of its environment, and it facilitates increased specialization of brain regions (Sowell et al. 1999; Rubia et al. 2000; Sowell et al. 2003).

Different parts of the brain follow different developmental courses. Gray matter structures generally follow an inverted "U" trajectory, whereas white matter increases during the first four decades. Structures related to processing information from the five senses – sight, sound, touch, taste, feel – mature early, within the first few months of life. The part of the brain involved in processing incoming speech (i.e., Wernicke's area) matures before the part involved in the production of speech (i.e., Broca's area). Areas involved in integrating activity from parts of the brain

involved in more primary processing mature later (i.e., association areas) and circuitry integrating input from the association areas mature later still (i.e., high association areas). To use a literary metaphor, if the brain components are like letters of the alphabet, maturation proceeds by combining the letters into words, then the words into sentences, the sentences into paragraphs, and so on. The high association areas (e.g., prefrontal cortex) synthesize the lower levels of organization into the higher (e.g., words to sentences; sentences to paragraphs) and continue to change dynamically well into the third decade of life.

It is the relatively late maturation of the prefrontal cortex that has garnered the most attention in judicial realms. Areas of the prefrontal cortex are part of circuitry involved in many of the traits and behaviors most relevant to the issues of informed consent and culpability: (1) controlling impulses; (2) imagining and weighing the consequences of different courses of action (i.e., counterfactuals); (3) delaying gratification; (4) judgment; and (5) long range planning. It is not that these capabilities are absent during the teen years, but they are not as good as they are going to get. Like other features of the adolescent brain, the late maturation of the prefrontal cortex is not inherently a liability. Across species ranging from corvids (e.g., crows, ravens, rooks, jays, magpies) to primates, protracted development and greater time of dependence upon caregivers is correlated with larger brains and greater cognitive capacities (Cluxton-Brock 1991; Holzhaider et al. 2010). The extremely protracted development of *Homo sapiens* allows our species to stay flexible to changing environmental challenges and might have been crucial to our survival.

The Changing Frontal/Limbic Balance

The late maturation of the prefrontal cortex (and other higher association areas) is in stark contrast to other brain structures more laden with hormone receptors that are ignited at the time of puberty. Sometimes referred to generically as “limbic,” these areas are located more centrally in the brain and are involved in circuitry of emotions, reward systems, and drives. The jolt to limbic circuitry at puberty triggers several changes important for our survival. Sex drive becomes prominent to perpetuate our species. Aggression increases to secure and protect resources. Social interactions, one of our most effective survival skills, take on heightened priority.

Compared with adults and younger children, adolescent decision-makers are particularly sensitive to environmental cues, are particularly motivated by rewarding stimuli, and struggle to manage their affective responses (Casey et al. 2008; Somerville et al. 2010). A neural mechanism underlying these tendencies is increased dopaminergic activity in the circuit that connects the amygdala (a limbic structure involved in emotional processing), the ventral striatum (a key component of the brain’s motivational circuitry), and the frontal lobes (Benes 1998; Cunningham et al. 2002; Somerville et al. 2010). The dopamine system is part of a behavioral activation network that alters incentive-motivated acts (Siever 2008; Wahlstrom et al. 2010). The dopaminergic system undergoes dramatic changes during adolescence including a sharp increase in the number of dopamine receptors.

This is consistent with functional MRI studies demonstrating larger activation in the nucleus accumbens, orbitofrontal cortex, and anterior cingulate cortex in response to rewarding stimuli (Wahlstrom et al. 2010).

Several theorists (e.g., Dahl 2001; Steinberg 2007; Ernst et al. 2009) have posited that adolescent risk-taking behavior is related to a temporal gap between this early dopaminergic surge and the “slow and steady” development of the cognitive control system. We can see evidence of this gap in studies designed to probe motivated behavior, delay of gratification, and peer influence. For example, a driving-simulator fMRI study found that when adolescents knew their driving was being observed by peers, they exhibited much greater activation in the reward circuitry (i.e., ventral striatum and orbitofrontal cortex) than when they completed the task without being watched by their peers (Chein et al. 2010). Further, reward activation in the presence of peers was strongly negatively correlated with their self-reported resistance to peer influence (Chein et al. 2010). Notably, adults engaged in the same task were not affected by being observed.

Like the dopamine system, the serotonin system also undergoes substantial changes during adolescence. In addition to its involvement in basal ganglia structures, serotonin also modulates the expression of aggression via actions in the orbital frontal cortex and anterior cingulate cortex. Selective serotonin reuptake inhibitors (SSRIs) may reduce impulsive aggression, and serotonin reduction is correlated with decreased learning of cooperation and a lowered perception of trustworthiness (Siever 2008).

Earlier models explaining behavior as the outcome of a “battle” between limbic and prefrontal cortex influences (e.g., id versus ego) have been replaced by a more nuanced view. In this paradigm, behavior emanates from a synergistic interaction; limbic input prioritizes and assigns meaning for the prefrontal cortex to integrate with past, present, and future considerations when initiating a course of action. Appreciating the interplay between limbic and frontal systems is imperative for understanding decision making during adolescence. If criminality and other risky behaviors were simply linked to prefrontal cortex maturation, one would expect children, not adolescents, to be the main offenders.

Connectivity

The increasing connectivity among disparate brain regions during youth development parallels increasing white matter as the brain progresses from infancy, to childhood, to adolescence, to adulthood. Unlike gray matter, white matter volumes increase more or less linearly throughout adolescence, providing the neural foundation for more complex, integrative cognitive processes (Paus et al. 2001; Anderson 2002). Increasing white matter volumes likely reflect a number of cellular processes (Paus 2010), but most notably, myelination. Myelin, a sheath of fatty cell material wrapped around neuronal axons, allows nerve impulses to travel throughout the brain more quickly and efficiently and helps to regulate the neural activity that facilitates information processing (Fields and Stevens-Graham 2002).

The electrical insulating properties of myelin allow for signals to travel at speeds up to 100× faster than for unmyelinated axons. Also, in myelinated axons, the ion pumps that fuel axonal activation only need to reset the ion gradients at nodes between sections of myelin, instead of along the entire expanse of the axons. This results in up to a 30-fold increase in the frequency with which a given neuron can transmit information. The combination of increased speed (100×) and quicker recovery time (30×) can yield a 3,000-fold increase in the amount of information transmitted per second. This non-trivial impact of myelin on the brain's ability to process information may underlie many of the aspects related to what the Supreme Court described as "immaturity, impetuosity, and failure to appreciate risks and consequences" (Luna and Sweeney 2004; Luna et al. 2010).

Empirical research linking brain anatomy and connectivity to adolescent behavior is beginning to accumulate. For example, individual differences in structural connectivity (i.e., organization of white matter tracts) have been linked to increased willingness to wait for a desired reward (Olson et al. 2009). Similarly, a handful of studies have linked individual differences in brain structure and/or activation to resistance to peer pressure (Grosbras et al. 2007; Paus et al. 2008). For example, in one study, 10-year olds who demonstrated high resistance to peer influence exhibited markedly more functionally integrated neural processes when processing nonverbal social cues than to same-aged children who demonstrated low resistance to peer influence (Grosbras et al. 2007).

Different Context, Different Person? Hot and Cold Cognition

Why, in some circumstances, do adolescents behave, decide, and reason like adults, and in other situations appear to have very limited maturity of judgment? Most developmental researchers agree that adult-like maturity is evident earlier in adolescence for intellectually driven decisions such as whether to seek contraceptives, or treatment for depression than for impulsive, peer-influenced decision-making situations (Steinberg et al. 2009). While long documented in the behavioral science literature (and lamented by parents of adolescents), evidence from structural and functional neuroimaging studies is now helping to provide the neural basis for this duality. The answer lies in the Achilles heel of adolescent maturity: socio-emotional context – particularly the distinction between "hot" and "cold" cognition.

Hot cognition refers to conditions of high emotional arousal, conflict, or peer pressure; this is often the case for the riskiest of adolescent behaviors (MacArthur Foundation Research Network on Adolescent Development and Juvenile Justice 2006). Cold cognition refers to non-stressed decision-making. The biology of "hot" versus "cold" cognition relates back to the interaction between prefrontal cortex and limbic structures. As described above, hot and cold cognition are supported by different neuronal circuits and have different developmental courses (Steinberg 2005). Hot cognitive situations shift motivational circuits into overdrive, leaving bare adolescents' most fundamental cognitive vulnerabilities. As functional connectivity increases between motivational and cognitive control circuits, great socio-emotional maturity, and the ability to reliability demonstrate it, can be expected (Luna et al. 2001).

Neuroethical Cautions and Pitfalls

The promise of a biological explanation for adolescent risk behavior has captured the attention of the media, parents, policymakers, and clinicians alike. There is still, however, much to be learned about how changes in brain structure and function relate to adolescent behavior. A few cautions are warranted.

First, it is important to avoid the ecological fallacy, i.e., using findings about groups of people to make predictions about a specific individual. This is akin to using average monthly rainfall data to predict the likelihood that it will rain on any one day. Neuroimaging studies often create a "composite brain" by combining data from several individuals in order to make inferences. Consequently, neuroscience research is well equipped to address group-level difference in brain structure and function (e.g., sex differences or differences between 20-year olds and 14-year olds). However, it is more complicated to use those brain composites to predict an individual's brain structure, function, or behavior. There have been increasing efforts to using imaging to link individual differences in brain development to behavior, and while these studies show promise, this work is still in its infancy (Dosenbach et al. 2010).

Second, while our judicial system has affirmed age 18 as key legal transition between childhood and adulthood, even at the group level, there is not a sharp transition between age 17 and 18, or "adolescent" to "adult" brain. Further, because the time course of development is highly variable from person to person, a person's chronological age alone may tell us relatively little about his or her level of neurological or behavioral maturity (Aronson 2007).

Third, given the enormous role of context in shaping brain and behavior, some caution is warranted when generalizing from research conducted in the laboratory. Behavior is a function of multiple interactive influences including experience, parenting, personality, culture, psychological well-being, and social relationships (Arnett 1992; Irwin Jr et al. 1992). It remains challenging to create ecologically valid scenarios in the laboratory that replicate the full array of factors that influence behavior generally and maturity of judgment specifically. Thus, an adolescent's performance on laboratory tasks may not reflect the level of maturity he or she would demonstrate in the real world.

The Power of Brain Science

It is tempting to view neuroimaging as means of bypassing the vagaries of human behavior in favor of a more tangible, objective link to an individual's intentions or capacities. Neuroscience wields enormous influence in our current social dialogue. Judges have sometimes been reluctant to allow brain images into the court room, for fear that their visual appeal and scientific aura may be overly persuasive (Hughes 2010). Although the most recent research suggests that brain images per se may be no more persuasive than other ways of presenting findings (e.g., Farah and Hook 2013), neuroscience more generally may exert disproportionate influence. Neuroscience research findings, even if logically irrelevant, increase an

explanation's persuasiveness (Weisberg et al. 2008). Pictures of the brain satisfy our basic need to see, hold, and manipulate concepts that are complex and diffuse, like maturity, development, and motivation. Consequently, brain scans are popularly seen as "hard" evidence, while behavioral science data are seen as subjective. In reality, behavioral scientists and neuroscientists are often telling the same story, but neuroscientists are much more likely to be heard.

Conclusion

The issues of informed consent and legal culpability illustrate the emerging neuroethical challenges and opportunities involved in using neuroscience to inform child and adolescent policy. The use of neuroimaging in policy, in particular, highlights that the complexities of human behavior transcend pictures of the brain. The ability see the brain and to watch it change from childhood to adulthood has been transformative; nonetheless, these pictures do not eliminate the uncertainties involved in balancing adolescents' need for protection and their growing autonomy. At any point in development, the brain reflects a lifetime of interaction among biology, behavior, and environment (Johnson et al. 2009; Johnson and Blum 2012). Similarly, at any moment in childhood and adolescence, the capacities for cognitive and socio-emotional maturity are a function of both characteristics of the brain and characteristics of the decision-making environment. These complexities are rife for misinterpretation, misapplication, and misunderstanding. Brain scientists, ethicists, and legal and medical professionals must balance the need to communicate scientific complexity and uncertainty, with the need to communicate clearly what is not in question; at the threshold to adulthood, the brain is changing, and a continuing policy dialogue about how best to accommodate, support, and understand those changes is essential.

Cross-References

▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
▶ Developmental Neuroethics
▶ Justice: A Neuroanthropological Account
▶ Mind, Brain, and Law: Issues at the Intersection of Neuroscience, Personal Identity, and the Legal System
▶ Neurolaw: Introduction

References

Aber, L., Atkins, M., et al. (2012). Brief of Amici Curiae in No. 10–9646 & No. 10–9647.

American College of Obstetricians and Gynecologists. (2009). *Tool kit for teen care*. Washington, DC: American College of Obstetricians and Gynecologists.

American Psychological Association. (1989). *Amicus curiae brief filed in U.S. Supreme Court in Ohio v. Akron Center for Reproductive Health, Inc., 497 U.S. 502 (1990) and Hodgson v. Minnesota, 497 U.S. 417 (1990)*. Retrieved November 27, 2012, from http://www.apa.org/about/offices/ogc/amicus/hodgson.aspx

American Psychological Association and the Missouri Psychological Association. (2004). *Amicus brief to the Supreme Court of the United States Supporting Christopher Simmons in Roper vs. Simmons* (Case No. 03–633).

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Neuropsychology, Development and Cognition, Section C: Child Neuropsychology, 8*(2), 71–82.

Arnett, J. J. (1992). Reckless behavior in adolescence: A developmental perspective. *Developmental Review, 12*(4), 339–373.

Aronson, J. (2007). Brain imaging, culpability and the juvenile death penalty. *Psychology, Public Policy, and Law, 13*(2), 115–142.

Benes, F. M. (1998). Brain development, VII: Human brain growth spans decades: Carol A. Tamminga, M.D., Editor. *American Journal of Psychiatry, 155*(11), 1489.

Braitenberg, V. (2001). Brain size and number of neurons: An exercise in synthetic neuroanatomy. *Journal of Computational Neuroscience, 10*(1), 71–77.

Casey, B. J., Getz, S., et al. (2008). The adolescent brain. *Developmental Review, 28*(1), 62–77.

Cauffman, E., & Steinberg, L. (2012). Emerging findings from research on adolescent development and juvenile justice. *Victims & Offenders, 7*(4), 428–449.

Chein, J., Albert, D., et al. (2010). Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry. *Developmental Science, 14*(2), F1–F10.

Cluxton-Brock, T. (1991). *The evolution of parental care*. Princeton: Princeton University Press.

Cunningham, M. G., Bhattacharyya, S., et al. (2002). Amygdalo-cortical sprouting continues into early adulthood: Implications for the development of normal and abnormal function during adolescence. *Journal of Comparative Neurology, 453*(2), 116–130.

Dahl, R. E. (2001). Affect regulation, brain development, and behavioral/emotional health in adolescence. *CNS Spectrums, 6*(1), 60–72.

Derish, M. T., & Heuvel, K. V. (2000). Mature minors should have the right to refuse life-sustaining medical treatment. *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics, 28*(2), 109–124.

Dosenbach, N. U. F., Nardos, B., et al. (2010). Prediction of individual brain maturity using fMRI. *Science, 329*(5997), 1358–1361.

Ernst, M., Romeo, R. D., et al. (2009). Neurobiology of the development of motivated behaviors in adolescence: A window into a neural systems model. *Pharmacology, Biochemistry, and Behavior, 93*(3), 199–211.

Farah, M. J., & Hook, C. J. (2013). The seductive allure of "Seductive Allure". *Perspectives on Psychological Science, 8*(1), 88–90.

Feld, B. (1993). Criminalizing the American juvenile court. In B. Tonry (Ed.), *Crime and justice: An annual review of research* (p. 17). Chicago: University of Chicago Press.

Fields, R. D., & Stevens-Graham, B. (2002). New insights into neuron-glia communication. *Science, 298*(5593), 556–562.

Ford, C., Bearman, P. S., et al. (1999). Foregone health care among adolescents. *JAMA, 282*(23), 2227–2234.

Ford, C., English, A., et al. (2004). Confidential health care for adoelscents: Position paper of the Society for Adolescent Medicine. *Journal of Adolescent Health, 35*, 160–167.

Giedd, J. N., Blumenthal, J., et al. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience, 2*(10), 861–863.

Grosbras, M. H., Jansen, M., et al. (2007). Neural mechanisms of resistance to peer influence in early adolescence. *The Journal of Neuroscience, 27*(30), 8040–8045.

Holzhaider, J. C., Hunt, G. R., et al. (2010). Social learning in New Caledonian crows. *Learning & Behavior, 38*(3), 206–219.

Hughes, V. (2010). Science in court: Head case. *Nature, 464*, 340–432.
Irwin, C. E. Jr., Millstein, S. G., et al. (1992). Risk-taking behaviors and biopsychosocial development during adolescence. In *Emotion, cognition, health, and development in children and adolescents* (pp. 75–102). Hillsdale: Lawrence Erlbaum.
James, T. (1960). The age of majority. *The American Journal of Legal History, 4*, 22–33.
Johnson, S. B., & Blum, R. W. (2012). Stress and the brain: How experiences and exposures across the life span shape health, development, and learning in adolescence. *Journal of Adolescent Health, 51*(2), S1–S2.
Johnson, S. B., Blum, R. W., et al. (2009). Adolescent maturity and the brain: The promise and pitfalls of neuroscience research in adolescent health policy. *The Journal of Adolescent Health, 45*(3), 216–221.
Luna, B., & Sweeney, J. A. (2004). The emergence of collaborative brain function: FMRI studies of the development of response inhibition. *Annals of the New York Academy of Sciences, 1021*, 296–309.
Luna, B., Thulborn, K. R., et al. (2001). Maturation of widely distributed brain function subserves cognitive development. *NeuroImage, 13*(5), 786–793.
Luna, B., Padmanabhan, A., et al. (2010). What has fMRI told us about the development of cognitive control through adolescence? *Brain and Cognition, 72*(1), 101–113.
MacArthur Foundation Research Network on Adolescent Development and Juvenile Justice. (2006). Issue Brief 3: Less guilty by reason of adolesence (September 21, 2006).
Olson, E. A., Collins, P. F., et al. (2009). White matter integrity predicts delay discounting behavior in 9- to 23-year-olds: A diffusion tensor imaging study. *Journal of Cognitive Neuroscience, 21*(7), 1406–1421.
Paus, T. (2010). Growth of white matter in the adolescent brain: Myelin or axon? *Brain and Cognition, 72*(1), 26–35.
Paus, T., Collins, D. L., et al. (2001). Maturation of white matter in the human brain: A review of magnetic resonance studies. *Brain Research Bulletin, 54*(3), 255–266.
Paus, T., Toro, R., et al. (2008). Morphological properties of the action-observation cortical network in adolescents with low and high resistance to peer influence. *Social Neuroscience, 3*(3–4), 303–316.
Robertson, S. (2008). *Age of consent laws in children and youth in history, Item #230*. Retrieved January 14, 2013, from http://chnm.gmu.edu/cyh/teaching-modules/230
Rubia, K., Overmeyer, S., et al. (2000). Functional frontalisation with age: Mapping neurodevelopmental trajectories with fMRI. *Neuroscience & Biobehavioral Reviews, 24*(1), 13–19.
Siever, L. J. (2008). Neurobiology of aggression and violence. *The American Journal of Psychiatry, 165*(4), 429–442.
Somerville, L. H., Jones, R. M., et al. (2010). A time of change: Behavioral and neural correlates of adolescent sensitivity to appetitive and aversive environmental cues. *Brain and Cognition, 72*(1), 124–133.
Sowell, E. R., Thompson, P. M., et al. (1999). In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nature Neuroscience, 2*(10), 859–861.
Sowell, E. R., Thompson, P. M., et al. (2001). Mapping continued brain growth and gray matter density reduction in dorsal frontal cortex: Inverse relationships during postadolescent brain maturation. *The Journal of Neuroscience, 21*(22), 8819–8829.
Sowell, E. R., Petersen, B. S., et al. (2003). Mapping cortical change across the human life span. *Nature Neuroscience, 6*(3), 309–315.
Spear, L. P. (2000). The adolescent brain and age-related behavioral manifestations. *Neuroscience & Biobehavioral Reviews, 24*(4), 417–463.
Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences, 9*(2), 69–74.
Steinberg, L. (2007). Risk-taking in adolescence: New perspectives from brain and behavioral science. *Current Directions in Psychological Science, 16*, 55–59.
Steinberg, L. (2009). Adolescent development and juvenile justice. *Annual Review of Clinical Psychology, 5*(1), 459–485.

Steinberg, L., & Scott, E. S. (2003). Less guilty by reason of adolescence: Developmental immaturity, diminished responsibility, and the juvenile death penalty. *American Psychologist, 58*(12), 1009–1018.

Steinberg, L., Cauffman, E., et al. (2009). Are adolescents less mature than adults?: Minors' access to abortion, the juvenile death penalty, and the alleged APA "flip-flop". *The American Psychologist, 64*(7), 583–594.

Thomas, N., & O'Kane, C. (1998). The ethics of participatory research with children. *Children & Society, 12*(5), 336–348.

United States Supreme Court. (1990). Hodgson vs. Minnesota, 497, US 417.

United States Supreme Court. (2005). Roper vs. Simmons, US Supreme Court.

United States Supreme Court. (2010). Graham vs. Florida, US Supreme Court.

United States Supreme Court. (2012). Miller vs. Alabama, 567.

Wahlstrom, D., White, T., & Luciana, M. (2010). Neurobehavioral evidence for changes in dopamine system activity during adolescence. *Neuroscience and Biobehavioral Reviews, 34*(5), 631–648.

Weisberg, D. S., Keil, F. C., et al. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477.

Neuroscience, Gender, and "Development To" and "From": The Example of Toy Preferences

110

Cordelia Fine

Contents

Abstract

"Development to" perspectives implicitly or explicitly assume that experience influences the individual's development "to" a genetically encoded phenotype. By contrast, "development from" perspectives assume no genetically pre-specified developmental pathway, but the co-construction of the phenotype from the complex and dynamic interaction between environmental stimuli, genotype, and the organization of the nervous system at each developmental phase. This chapter examines the "brain organization" account of sex differences in toy preferences in light of challenges to the "development to" perspective, of which the brain organization account is an example. It is argued that there are significant methodological and conceptual issues, and empirical uncertainties, regarding each of four categories of evidence commonly cited as support for the brain organization account. The scientific and

C. Fine
Melbourne School of Psychological Sciences & Melbourne Business School & Centre for Ethical Leadership, University of Melbourne, Carlton, VIC, Australia
e-mail: c.fine@mbs.edu

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_151,

ethical need for research from a "development from" perspective for future investigations of this politically important and socially sensitive scientific question is discussed.

Introduction

Persistent sex differences in social roles, occupations, and occupational success – even in progressive twenty-first century societies – is a phenomenon requiring explanation. As Eagly and Wood (2013) have recently noted, research within behavioral science that ultimately seeks to answer this question has mostly taken place in parallel streams. On the one hand, there are researchers interested in the influence of social factors (such as gender socialization and sex-based discrimination), while others investigate the contribution of biological factors (such as the effect of hormonal and brain differences between the sexes). The conceptual framework for the latter group of researchers is often (either explicitly or implicitly) the prominent and influential "brain organization" theory (for brief overview, see Breedlove et al. 1999; Hines 2010). During gestation, the gene-directed development of testes in the male fetus results in a surge of gonadal fetal testosterone (fT), and this directs the development of male genitalia. In humans, this "critical period" occurs during approximately weeks 8–24 of gestation (Reyes et al. 1973). Brain organization theory, first proposed by Phoenix and colleagues (Phoenix et al. 1959), holds that a second effect of this surge of fT is to permanently "organize" a male brain that produces male behavior (in some cases, after these brain structures are activated by circulating sex hormones in pubescence and adulthood).

Brain organization theory was originally proposed to explain sexually differentiated behavior, particularly behavior tied to reproduction, in nonhuman animals, but it has since been proposed that the organizational effects of fetal testosterone on brain development contribute to human sex differences in sexuality, gender identity, and gender-typed interests (e.g., Hines 2010, 2011; for comprehensive review and critique, see Jordan-Young 2010). Proponents of such brain organization accounts of course acknowledge that social experiences contribute to human sex differences. However, social experience is often implicitly or explicitly represented as playing a merely influential, amplifying, or even interfering role in development, rather than an integral one as co-author of the developing phenotype. Moore (2002) has provided a helpful articulation of the distinction between these differing perspectives, referring to them as "development to" and "development from" approaches, respectively. In the "development to" perspective, there is an underlying assumption that experience merely influences the individual's progress "to" a genetically encoded phenotype. By contrast, according to a "development from" perspective, there is no pre-specified developmental pathway. Rather, every developmental step is constructed from the complex and dynamic interaction between environmental stimuli (including social experiences), genotype, and the organization of the nervous system in a particular developmental phase.

A "development from" approach, with bidirectional genetic, hormonal, neuronal, behavioral, environmental, and cultural influences, is supported by research across each level of analysis of behavior (Li 2003; Lickliter and Honeycutt 2003). This has important implications for neurobiological investigations of sex differences: Gender, as a powerful and pervasive social phenomenon, has material effects on the body and brain (e.g., Fausto-Sterling 2005; Kaiser 2012). One recent example is provided by a large-scale longitudinal study, which found that fatherhood reduced testosterone levels in men. This reduction was greater in fathers who spent more time physical caring for their young offspring (Gettler et al. 2011). The influence on endocrine state of the social construction of gender roles for fathers, in terms of expected contribution to parenting, is also indicated by a comparison of two neighboring cultural groups in Tanzania, which found lower testosterone levels among fathers from the population in which paternal care was the cultural norm, than in fathers from the other group in which paternal care was typically absent (Muller et al. 2009).

Also in line with a "development from" perspective is the growing evidence for neuronal plasticity throughout development. Neuronal plasticity refers to phenomena whereby neuronal characteristics are responsive to external, including social, experiences, resulting in changes such as in neuronal responsiveness, synaptic connectivity, dendritic branching, gene-expression, and gray and white matter volumes (e.g., Draganski et al. 2004; Edelmann and Auger 2011; Fields 2010; Haier et al. 1992; Jäncke et al. 2001; Maguire et al. 2000). Thus, while clearly the brain is not infinitely malleable, neural circuitry develops through, and is altered by, experience (Westermann et al. 2007). Accordingly, the social phenomenon of gender – in which behavior and activities are influenced by stereotypes and norms that are variable across time and place – becomes "part of our cerebral biology" (Kaiser et al. 2009, p. 57).

Importantly, a "development from" perspective makes clear the error of conceptualizing variables such as hormonal level, hormonal effects on brain or behavior, or neural function or structure, as pure biological variables. Rather, they are intertwined with the individual's life-history and current social context, and sex difference research that proceeds without this insight runs the risk of being misleading and/or uninformative. As Moore (2002, p. 65) pointed out in relation to research conducted within the brain organization framework, research strategies that work from a "development to" perspective, and therefore observe only early hormones and later behavioral outcomes, leave "lots of unexplored territory and many possible pathways, perhaps convoluted ones, from the early hormones and end points of interest." In other words, such strategies neglect to investigate the complex, dynamic process of development itself. Moore's work, demonstrating the unexpectedly complex effects of early testosterone on sex-differentiated brain stem characteristics and sexual behavior in rats, has provided an increasingly well-known (although long ignored, see Kaplan and Rogers 2003) example of the problematic nature of a "development to" research approach. Mother rats are attracted to odor cues from the higher levels of testosterone in the urine of males and, therefore, lick and groom male pups more than they do female pups.

Remarkably, this differential maternal treatment of males and females contributes to low-level sexually differentiated brain structure and sexual behavior (Moore 1984; Moore et al. 1992). According to the standard brain organization account, these brain and behavioral differences would be attributed solely to the direct action of early testosterone on the brain.

In light of these challenges to the "development to" perspective implicit in the brain organization account, this chapter examines an important and influential hypothesis derived from the brain organization account. This is the claim that sex differences in fetal testosterone, via permanent effects on brain structure, create inherent differences in sex-typed interests that are reflected in male/female differences in toy preferences in childhood. This behavioral difference is among the most substantial observed in childhood, and exceeds those found in cognition and personality (Hines 2010). Four lines of research are repeatedly put forward as evidence for a brain organization account of sex differences in toy preferences. First, it is argued that sex differences in visual interest in a social versus mechanical stimulus exist even in the first few days of life, prior to either the development of gender identity or exposure to significant gender socialization processes. Second, females with the genetic condition of congenital adrenal hyperplasia (CAH), who are exposed to atypically high levels of fetal testosterone during gestation, show more male-typical toy and activity preferences than do non-affected female controls. Third, it is claimed that sex differences in toy preferences similar to those observed in human children are also found in rhesus and vervet monkeys. Fourth, correlations have sometimes been observed between markers of fetal testosterone levels (taken during the critical period) and later gendered play preferences. These four lines of evidence are regularly presented, both in the scientific literature as well as in popular accounts, as showing that the brain organization account of sex differences in toy preferences is beyond reasonable doubt (e.g., Hines 2011; Hoff Sommers 2012; Orenstein 2011; Saad 2012; Wong et al. 2012).

However, the discussion above raises a priori reasons to suppose that such a conclusion may be premature; and, in fact, substantial methodological and conceptual criticisms have been made of each of the lines of evidence marshaled in support of the hypothesis. The following sections briefly summarize findings, and review and expand critiques of these studies' methods and conclusions. The final section argues, on both scientific and ethical grounds, a need to better take these criticisms into account, and deploy a "development from" perspective in future research.

The Brain Organization (Masculinization) Account of Sex Differences in Toy Preferences

Many toys enable children to role-play adult activities, and some of these activities and their associated toys are more strongly linked with one sex than with the other (e.g., dolls and tea-sets vs. trucks and guns). Although many behavioral differences between the sexes are modest both in children and adults (Hyde 2005),

from 3 years of age there are substantial sex differences in toy preferences. For example, in a typical lab-based observational study, children were offered a collection of female-typical, male-typical, and neutral toys to play with. Girls spent about 60 % of their playing time with female-typical toys (a set of dishes, a Barbie doll with clothing and accessories, a rag doll with accessories and a cosmetics kit), while boys spent only 6 % of their time on those toys. By contrast, the boys spent 70 % of their time with male-typical toys (a car, a fire-truck, a Lego airplane or Lincoln Logs construction toy, a tool set, a helicopter and a gun), compared with only 13 % for girls (Pasterski et al. 2005). Girls and boys spent similar amounts of time playing with the neutral toys (a puzzle, a board game, books, crayons, and a sketchpad). Parents also report sex differences in questionnaires about their children’s sex-typed childhood activities and interests (e.g., Hines et al. 2004).

Why do these sex differences in toy and activity preferences exist? Self-socialization perspectives emphasize the salience and importance of gender in the social world (Bem 1983), and the motivating effect of this on children, who play an active role in their own gender development once they become aware of their gender identity at about 2 years of age (e.g., Arthur et al. 2008; Bigler and Liben 2007; Martin and Halverson 1981). The salience and functional importance of gender is also a component of social learning perspectives on toy preferences, although these emphasize instead the role of others (such as caregivers) in modeling, channeling, and reinforcing stereotype-consistent behavior (e.g., Bussey and Bandura 1999). These accounts therefore anticipate the appearance of sex differences even prior to the development of gender identity.

However, the brain organization account makes the additional proposal that innate brain differences, arising from sex differences in exposure to fT, contribute significantly to gendered toy preferences. Sex differences in prenatal testosterone levels are suggested to provide the “seeds” of later male/female differences in toy preferences, with “nurture” progressively recruited in ways that amplify these initial psychological biases (Alexander and Wilcox 2012; Baron-Cohen 2007; Berenbaum and Resnick 2007). As to what those psychological biases might be, Alexander (2003) suggested that males are born predisposed to be attracted to movement (since this would have advantaged them in developing hunting skills in prehuman and early humans societies), while females are born predisposed to be attracted to reddish-pink colors and rounded forms evocative of infants (since this would have advantaged them in developing infant nurturance skills). Related suggestions are that males might be more drawn to objects that allow for active play (see Alexander and Saenz 2012), or that they have “[i]nnate predispositions for perceptual attributes or motor affordances of objects” that bias them toward objects or activities that allow for propulsive movement (Benenson et al. 2011, p. 263). A second influential proposal, the Systemizing/Empathizing hypothesis, is that higher levels of fT predispose the (typically) male baby toward understanding and building rule-driven, input-function-output systems. By contrast, lower levels of fT predispose (typically) female infants to attend to empathy-related stimuli; namely,

people (Baron-Cohen 2003). As noted earlier, four lines of research are repeatedly put forward as evidence for a brain organization account of sex differences in toy preferences, and these are each now discussed in turn.

The Newborn Study

While a number of studies have looked for sex differences in toy preferences in infants, intended or unintended gender socialization processes, such as caregiver responses and familiarity effects, could potentially underlie any differences observed. To exclude this as a possibility, a much cited study compared neonates' looking time at a live face versus a mobile (Connellan et al. 2000). These stimuli were chosen to reflect interest in biological/social motion versus mechanical motion. Male and female babies both spent approximately half of the total presentation time looking at the face, which was that of the first author. However, males looked longer at the mobile than did females (52 % of presentation time vs. 41 % for females) and females looked longer at the face than at the mobile. These findings have been interpreted both by the study authors and others as evidence for "innate" sex differences in psychological interests.

However, serious concerns have been raised over the considerable methodological flaws of this study (Nash and Grossi 2007). These include the many differences between the stimuli (any of which could have been responsible for the observed differences), and nonstandard procedures for measuring looking time preference (such as serial rather than simultaneous presentation) and, in particular, the scope for experimenter expectancy effects. The first author was both the face stimulus and controlled the movement of the mobile, yet inadequate efforts were made to ensure that the experimenter was blind to the baby's sex. These are serious methodological shortcomings; moreover, the study has never been replicated. Indeed, a recent study of 4–5 month-old infants, that used a number of different face versus object stimuli, found that both girls and boys preferred faces (Escudero et al. 2013). In addition, no evidence is provided that a newborn's visual preference in this experiment anticipates his/her future abilities and interests: It is an assumption that is "essentially unargued for" and "questionable at best." (Levy 2004, p. 322; see also Nash and Grossi 2007). Indeed, it seems to implicitly assume that newborn visual preference is an early indicator of a future biologically pre-specified developmental outcome.

Toy Preferences in Females with CAH

The toy and activity preferences of females with CAH are of considerable interest to researchers, since they provide a group in which high levels of fT exposure are separated from social rearing as a boy. Studies investigating the toy preferences of females with CAH typically use parental and retrospective self-report questionnaires, and lab-based observational studies. Preferences are compared with those of boys and unaffected female relative controls. For example, studies have compared

scores on the Pre-School Activities Inventory (PSAI), which measures interest in toys and activities, and display of characteristics, that are differentially observed in boys and girls (Golombok and Rust 1993). Both questionnaire and observational studies have consistently found that females with CAH show a stronger preference for male-typical toys and activities, and less interest in female-typical ones, compared with unaffected female relative controls (e.g., Berenbaum and Hines 1992; Hines et al. 2004; Nordenström et al. 2002). For example, Pasterski et al. (2005) found that girls with CAH spent only 21 % of time playing with female-typical toys (compared to 61 % for unaffected girls), and 44 % of time with male-typical toys (compared to 13 %).

These findings are often regarded as providing definitive support for a brain organization account of toy preferences. However, these studies do not directly test its predictions. Clearly, the most appropriate way to test such hypotheses would be for researchers to categorize (or create) toys on the basis of the presence or absence of the features thought to be critical: object features such as movement, color, and form (Alexander 2003); affordance for active play (Alexander and Saenz 2012); or stimuli that represent rule-driven, input-function-output systems versus empathy-related stimuli (Baron-Cohen 2003). Instead, stimuli sets are created on the basis of their popularity with males versus females, a strategy criticized decades ago by Bleier (1986, p. 150) for its presumption that culturally defined masculinity is “as objective and innate a human feature as height and eye color.” Thus, when researchers observed that an assumedly “male-typical” toy (the Lincoln Log construction toy) was very popular with control girls, it was eliminated from the male-typical set (Pasterski et al. 2005). Clearly, if it had been chosen a priori on the basis of the presence of features thought to be intrinsically appealing to a masculinized brain, this would have instead constituted counter-evidence to the brain organization account. This approach is problematic, because the features supposedly attractive to a masculinized (non-masculinized) brain are neither exclusive to, nor always present in, male-typical (female-typical) toys. Toy vehicles can be moved, but so too can toy vacuum cleaners, prams, and pull-along toys, none of which are particularly associated with boys. Guns and construction toys do not afford movement more than, say, tea-sets. Cosmetics invite systemizing, since they involve the transformation of an input (the “before” state) into an output (the desired “after” state) via a function (the application of cosmetics). In addition, neither cosmetics nor jewelry are necessarily associated with reddish-pink colors, round features, or empathizing. Stuffed animals, by contrast, have some of these features, yet are rarely used as female-typical toys and are sometimes instead categorized as neutral toys.

Furthermore, neutral toys often arguably have attributes that should be differentially attractive to males and females, according to the Empathizing/Systemizing account (Fine 2010). For example, puzzles, board games, and books are frequently used as “neutral” toys. However, puzzles and board games are arguably “systemizing” activities, and are indeed referred to in a questionnaire designed to measure systemizing tendencies in children, the SQ-Child (Auyeung et al. 2006). In addition, books could arguably be categorized as an empathizing toy, at least in cases

where the book presents characters and their thoughts and emotions (either in text or illustration, depending on the age group under study). Similar issues arise when considering the use of the PSAI. Notably, the questionnaire was not developed to test brain organization accounts of play preferences, but rather to assess gender role behavior in pre-school children. It contains numerous items assessing behaviors that are presumably outside the scope of proposed effects of fT on the intrinsic value of object properties: e.g., interest in jewelry and pretty things, pretending to be a female character, avoidance of getting dirty, and dressing up in girlish clothes.

Thus, the current research approach makes it impossible to distinguish a brain "masculinization" explanation of findings, in which particular characteristics of toys are intrinsically more appealing to boys and females with CAH, from the alternative possibility that females with CAH are less attracted to whatever happens to be culturally ascribed to females and/or more attracted by a cultural ascription with males. This is an importantly different proposition, and there are good reasons to take this alternative account seriously. Gender identity in females with CAH is generally unremarkably female, but nonetheless differs modestly to that of female controls, with slightly more male identification and greater expression of dissatisfaction and unhappiness with a female gender identity (e.g., Berenbaum and Bailey 2003; Meyer-Bahlburg et al. 2006). Moreover, Jordan-Young (2010, 2012) has argued that, in their focus on prenatal hormone exposure, researchers have overlooked other variables also affected by the condition that plausibly influence psychosexual development. These include intensive medical and psychiatric intervention arising from atypical or masculinized genitalia, other physical effects of the condition inconsistent with cultural ideals of feminine attractiveness (such as hirsute appearance and short, heavy stature), and the priming of expectations of masculinity in parents, the girls themselves, and others. To date, the possible role of these other factors in the development of masculinized toy preferences has scarcely been investigated. Research in this direction is currently limited to questionnaires or observations of parental attitudes and behaviors regarding sex-typical and atypical play, the findings from which have been inconsistent (Berenbaum and Hines 1992; Pasterski et al. 2005; Wong et al. 2012), and represent only very early first steps in adequately acknowledging the physiological and psychological sequelae of the condition, as well as understanding how labeling, priming, and expectation effects arising from a diagnosis of CAH might, in complex and iterative ways, affect psychosexual development (see Jordan-Young 2012).

Sex Differences in Toy Preferences in Monkeys

Two studies of toy preferences in monkeys are often cited as support for the idea of "inborn" sex differences in predispositions toward different toy types. The first, an observational study of vervet monkeys' toy play behavior, compared contact time with male-typical toys (a ball and police car), female-typical toys (a toy pan and a doll), and neutral toys (a picture book and a stuffed dog), presented serially to groups of vervets (Alexander and Hines 2002). (As Jordan-Young (2010) has noted,

this procedure meant that any one individual vervet's choices were dependent on what other vervets were already playing with.) Between-sex contrasts showed greater male interest in the male-typical toys, and greater female interest in the female-typical toys. The sexes showed equal interest in the neutral toys. Within-sex contrasts found only that females had greater percentage contact with female-typical toys than with male-typical toys. A second study with rhesus monkeys compared interaction (using two dependent variables, total frequency and total duration of contact) with wheeled toys versus stuffed toys (Hassett et al. 2008). Between-sex contrasts found that males and females were equally interested in the wheeled toys. Males and females also spent a similar duration of time with the stuffed toys, but females had a greater total frequency of interaction with these toys. Within-sex contrast revealed that males preferred wheeled toys over stuffed toys, while females showed no preference.

As with the studies of females with CAH, interpretation of these two studies is complicated by the non-hypothesis-driven fashion in which toys have sometimes chosen by researchers. In particular, this approach has enabled male-typical, female-typical, and neutral toys to be categorized differently across studies, with the unintended effect of making findings appear more consistent than they actually are (see Jordan-Young 2010). For example, Servin and colleagues (Servin et al. 2003) classified a ball as a "neutral" toy, and it was the most popular toy (when presented with a choice between a car, ball, and doll) among (control) female girls. However, balls were categorized as a male-typical toy in the vervet monkey study. Similarly, a stuffed animal was a neutral toy in the vervet study, but the sole type of feminine toy in the rhesus monkey study. Importantly, male vervet monkeys played more with the stuffed dog (their favorite toy as a group) than with the car – this is in direct contradiction with the main finding of the rhesus monkey study.

An additional issue with the choice of toys for the monkey studies is that such stimuli are unlikely to hold the same meaning to monkeys as they do to human children, and the "affordances" monkeys might perceive in them are more assumed than proven. For example, as Jordan-Young (2010, p. 236) points out, "[h]ow does a vervet know that the purpose of a cooking pot is not to bang it, throw it, or use it to whack another vervet?" Jordan-Young has noted that although that study's findings were accompanied by a photo of a male vervet rolling the toy car along the ground, and a female vervet cradling the doll, the frequencies of such behaviors in each sex were not reported. Similarly, Hassett and colleagues chose stuffed animals versus wheeled toys to elicit evidence of different activity preferences, but although data were collected on the specific kinds of behaviors directed toward toys, these were not reported. It is therefore unknown whether, for example, stuffed toys tended to be nurtured, bitten, or thrown (indeed, one trial had to be terminated early when a stuffed toy "was torn into multiple pieces"), or whether play with wheeled toys was more active or involved more object movement than play with stuffed animals. A recent study with young children found that play with female-typical toys was as active as play with male-typical ones (Alexander and Saenz 2012). Furthermore, as Ah-King (2009) has noted, since vervet monkeys

are tree-dwelling vegetarians, it is unclear why males in particular should show a predisposition for the development of hunting skills. Rather, the ability to navigate in space would be necessary for survival in both sexes.

One final point of criticism targets the assumption that sex differences in monkeys' toy preferences cannot be attributed to socialization processes (see Fine 2010). Like humans, primate societies have norms regarding sex roles (such as who gets food, cares for infants, etc.), and these norms can differ across, or even within, species (Burton 1977). For example, male involvement in infant rearing can range from absent to highly involved, even within the same species (Itani 1959; Burton 1992). Burton (1992, p. 45) reported extensive and lengthy male care of young in a Gibraltar troop of macaque species, with young females "kept away from infants so that young males may learn their role." She also observed imitation of infant care by the head male, by male subadults only, who then themselves became involved in infant care (Burton 1972). Interestingly, the behavior of male and female monkeys toward infants only starts to diverge at about 2–3 years of age (Mason 2002) and manipulations of fT exposure (both blocking in males, and increased exposure in females) have no effect on subsequent interest in infants (Herman et al. 2003). These findings indicate non-determination of roles by hormones, a significant role for social learning of sex roles, and challenge the assumption that sex differences in monkeys in play with infant-like toys, for example, must reflect "pure biology," absent the influence of socialization.

Associations Between Markers of Fetal Testosterone and Toy Preferences

A fourth category of studies regularly referred to are those that look for correlations between fT exposure and later gendered play preferences in childhood. The advantage of these studies is that they are based on nonclinical samples (although, in the case of populations who are sampled from mothers undergoing amniocentesis, they are not necessarily representative). Various markers of fT have been used (since ethically it is not possible to sample blood from the fetus unless medically indicated): Amniotic testosterone (aT) is sampled from the amniotic fluid during the procedure of amniocentesis; maternal testosterone (mT) is sampled from the mother's blood; and maternal sex hormone-binding globulin (SHBG), which limits T's functional effectiveness by binding with it, has been used as an inverse proxy for levels of unbound, functionally effective T.

To date, four studies have related markers of fT to later toy preferences (recently summarized in Grossi and Fine 2012, see Table 4.1). The first study assessed behavior using the PSAI and used both mT and maternal SHBG levels as proxies for fT exposure (Hines et al. 2002). In girls only, higher levels of mT (but not maternal SHBG) were associated with more masculine scores on the PSAI. The effect size was very small, explaining only two percent of the variance in score, and no other relationships were significant. (The possibility that mothers with higher vs. lower T levels might create different social experiences that influence their

daughters’ gendered preferences does not appear to have been considered.) Subsequently, Knickmeyer et al. (2005) looked for a relationship between aT and sex-typical play in 4 and 5 year old children, as measured by a questionnaire. No relationship with aT was found in either sex, or in both sexes together. Van de Beek and colleagues explored relationships between mT, aT, estradiol, and progesterone levels and observed play behavior in 13-month-old infants (Van de Beek et al. 2009). They found no relationships with aT, mT, or estradiol. Surprisingly, higher levels of amniotic progesterone were associated with a stronger preference for male-typical toys. Finally, in contrast with these mostly negative findings, Auyeung and colleagues, with a larger sample size, found correlations in both sexes, individually as well as pooled, between aT and PSAI score (Auyeung et al. 2009). It is unclear why this study, with a sample approximately one-third of the size of Hines et al. (2002), found a relation with males that was absent in the earlier study (see Jordan-Young 2010), and whether future work will support these positive relations.

A fifth study, an investigation of Baron-Cohen’s (2003) Systemizing/Empathizing hypothesis, looked for the predicted relations between aT and tendency to prefer systemizing activities in children, using a parental-report questionnaire, the SQ-Child (Auyeung et al. 2006). aT was significantly associated with SQ-Child score (which was greater for boys than for girls), both across the whole sample and for boys and girls separately. However, in addition to the subjectivity of parental report as opposed to observed behavior, only a small number of the items appear to reflect “the drive to analyze or construct systems,” with many items instead appearing to tap into a drive for order, routine, or arrangement of objects (Auyeung et al. 2006, p. S124; see Fine 2010; Grossi and Fine 2012).

One critical concern with these studies is that there is currently no satisfactory evidence that either aT or mT is related to actual fT exposure. In their review of this issue, van de Beek et al. suggested aT as the best index of fT exposure, but acknowledged the lack of knowledge regarding the relationship between levels of aT (the main source of which is fetal urine) and levels in the fetal blood (van de Beek et al. 2004). Indeed, one study that measured mT, aT, and fT between 15- and 23-weeks of gestation found no correlations between the three measures (Rodeck et al. 1985). A more recent clinical study did find that fT correlated with mT (Gitau et al. 2005). However, mT levels are not higher in women pregnant with boys than in those pregnant with girls (Hines et al. 2002; Rodeck et al. 1985), which suggests that “maternal serum androgen levels are not a clear reflection of the actual exposure of the fetus to these hormones” (van de Beek et al. 2004, p. 664). That markers of fT may not correlate with actual fT exposure is of considerable concern in terms of interpretation of findings (Fine 2010).

Conclusions and Future Directions

As noted earlier, a brain organization account of sex differences in toy preferences is regularly presented as though it were beyond reasonable doubt. Yet as this chapter has shown, there are significant methodological and conceptual issues,

and empirical uncertainties, surrounding each of the four categories of evidence. There is considerable tension between the brain organization account and the rejection of "development to" models within developmental science, and Jordan-Young (2010) has comprehensively documented the empirical inconsistencies and contradictions of the data supposedly supporting brain organization theory as applied to humans (see also Fine 2010; Grossi and Fine 2012).

Moreover, more generally, the scientific assumptions implicit in "development to" based accounts – that brain circuitry is largely fixed by a genetic blueprint, that there is unidirectional, causal pathway from genes to behavior via hormones and brains, and that evolution has left us with brains and mental processes strongly reminiscent of our Paleolithic ancestors – have been widely rejected following conceptual and empirical upheavals in the relevant scientific fields (see Fine et al. 2013). A "development from" perspective is more consistent with contemporary perspectives that humans have evolved an adaptively plastic brain that is responsive to environmental conditions and experiences, and the modulation of endocrine function by those experiential factors contributes to that plasticity (for relevant reviews, see Brown et al. 2011; Lickliter and Honeycutt 2003; May 2011; van Anders and Watson 2006). Together with evidence of the considerable variation, across time and place, in gender roles (see Wood and Eagly 2013), the need to question implicit assumptions – such as that current, Western categorizations of toys as masculine and feminine correspond precisely with innate predispositions, or that social learning can be overlooked when considering the behavior of nonhuman monkeys – becomes more obvious. The need to question implicit or explicit "development to" assumptions will be no less important as researchers attempt to relate fetal endocrine exposure to later brain states (e.g., Lombardo et al. 2012).

Children's play worlds are society writ small, and scientific accounts of how and why sex differences in play preferences develop therefore have important political ramifications. Continuing to ignore critiques of the research not only has scientific implications, but also political and ethical ramifications in terms of which groups benefit from what knowledge is produced, as well as from what knowledge is not produced (Haslam and McGarty 2001). Since power hierarchy and inequalities are embedded in gender as a social system, a scientific claim that presents gendered preferences as to some extent "innate" is not politically or socially inert. Brain organization accounts propose that prenatal hormones provide an initial "seed." This initial biologically based "seed" then recruits experience, as the child seeks out the kinds of toys and activities s/he finds most rewarding (although it is generally acknowledged that this is amplified by gender socialization processes). Thus, Baron-Cohen (Baron-Cohen 2007, p. 169) refers to socialization factors "amplifying" what is innately specified, and argues that "we should not expect the sex ratio in occupations such as math or physics to ever be 50–50 if we leave the workplace to simply reflect the number of applicants of each sex who are drawn to such fields." Berenbaum and Resnick (2007) describe hormonal influences as furnishing the "seeds of career choices" (p. 147), and "propose that sex-related career choices and outcomes arise through the mediating and moderating effects of socialization on sex-hormone-influenced individual differences in behavioral development." (p. 148).

Similarly, Alexander and Wilcox (2012, pp. 400–401) refer to hormonally produced “sex-linked dispositions that represent ‘seeds’ of later behavior,” and suggest that sex differences may be smaller in infancy than later in development “when expressed behavior presumably reflects the further influence of experiential factors.” Popular accounts of the social implications of the data often similarly subscribe to the compounding sex-segregation of interests, as “nature” recruits “nurture” (e.g., Hoff Sommers 2012). Importantly, since the “development to” perspective conceives of a unidirectional causal pathway from genes to hormones to brain to behavior, the biological is seen as causally primary in the developmental pathway. This privileging of the biological as somehow more “real” than social contributions is illustrated by the recent comment, by a leading researcher from the brain organization perspective, that research into sex differences in toy preferences “reveals both how humans develop, and how societal pressures act upon children.” (Hines 2013). That is, there is “real” (i.e., biologically based) development that social experiences then merely acts upon.

Brain organization accounts therefore see an original, “essential” difference between the sexes, which is then amplified by experience in a developmental cascade. Such “essentialist” views are associated with increased gender stereotyping, self-stereotyping, stereotype threat, and comfort with the gender status quo (see Fine 2012; see ▶ Chap. 91, “Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform” in this volume). While scientists may prefer to think that political and ethical values lie outside their domain of consideration, as the foregoing discussions indicate, such values are implicitly at work in the research questions that are asked, the rigor of the methodologies chosen, the background assumptions made, the emphasis on certain findings over others, and assessments of the uncertainty that is considered tolerable in order for a particular conclusion to be drawn (Douglas 2008; Haslam and McGarty 2001). Importantly, the future research directions that naturally arise out of the critiques presented here could all potentially produce knowledge that challenge an essentialist account of sex differences in toy preferences. For example, hypothesis-driven selection of toys based on the presence or absence of supposedly critical features (rather than cultural association with males vs. females) could potentially produce data that strongly challenge the brain organization account in a way that the current research approach cannot. It is noteworthy that such studies have never yet been conducted, despite Bleier’s critique of the standard approach nearly 30 years ago (Bleier 1986). Interestingly, the few recent studies that have investigated whether particular features of toys differentially appeal to males and females have not supported brain organization account proposals that females are drawn to pinkish-reddish colors (Jadva et al. 2010) or that males are drawn to toys that afford active play (Alexander and Saenz 2012), although Benenson et al. (2011) found sex differences in imitation of propulsive action in 6–9 month-old infants. Similarly, nonhuman primate research, building on recent findings from the social learning literature that these animals show discrimination in who they learn from (Mondragón-Ceballos et al. 2010; van de Waal et al. 2010), could seek to answer the question posed by Hines and Alexander (2008, p. 479): “if some animals of one sex could be trained to use a particular object, would others of that sex model

them?" Together with comparisons between groups with different sex role norms, again, such data could potentially destabilize the conclusion that human sex differences in toy preferences are "innate" and inevitable. Finally, in contrast with a "development to" perspective, a "development from" perspective allows for the possibility that, as is observed in animal hormonal studies, "an early push in a certain direction can be *either enhanced or entirely eliminated* by subsequent experience, such that development from that point forward would proceed as though the early hormone exposure had never happened." (Jordan-Young 2010, p. 288, emphasis in original). Greater attention needs to be paid to the complex and dynamic process of development itself when it comes to toy preferences (for discussion, see Fausto-Sterling et al. 2012).

In summary, for both scientific and social reasons, researchers need to incorporate a "development from" perspective that brings a conceptually sophisticated understanding of both development and gender to this politically important and sensitive scientific question.

Cross-References

▶ A Curious Coincidence: Critical Race Theory and Cognitive Neuroscience
▶ Developmental Neuroethics
▶ Feminist Ethics and Neuroethics
▶ Feminist Neuroethics: Introduction
▶ Feminist Philosophy of Science and Neuroethics
▶ Neuroethics and Identity
▶ Neuroethics of Neurodiversity
▶ Sex and Power: Why Sex/Gender Neuroscience Should Motivate Statistical Reform
▶ Toward a Neuroanthropology of Ethics: Introduction

References

Ah-King, M. (2009). Toy story: En vetenskaplig kritik av forskningom apors leksakspreferenser. *Tidskrift för genusvetenskap, 1*, 45–63.

Alexander, G. (2003). An evolutionary perspective of sex-typed toy preferences: Pink, blue, and the brain. *Archives of Sexual Behavior, 32*(1), 7–14.

Alexander, G., & Hines, M. (2002). Sex differences in response to children's toys in nonhuman primates (*Cercopithecus aethiops sabaeus*). *Evolution and Human Behavior, 23*, 467–479.

Alexander, G., & Saenz, J. (2012). Early androgens, activity levels and toy choices of children in the second year of life. *Hormones and Behavior, 62*, 500–504.

Alexander, G., & Wilcox, T. (2012). Sex differences in early infancy. *Child Development Perspectives, 6*(4), 400–406.

Arthur, A., Bigler, R., Liben, L., Gelman, S., & Ruble, D. (2008). Gender stereotyping and prejudice in young children: A developmental intergroup perspective. In S. Levy & M. Killen (Eds.), *Intergroup attitudes and relations in childhood through adulthood* (pp. 66–86). Oxford: OUP.

Auyeung, B., Baron-Cohen, S., Chapman, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Foetal testosterone and the child systemizing quotient. *European Journal of Endocrinology, 155*, S123–S130.

Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., Hackett, G., et al. (2009). Fetal testosterone predicts sexually differentiated childhood behaviour in girls and in boys. *Psychological Science, 20*(2), 144–148.

Baron-Cohen, S. (2003). *The essential difference: Men, women and the extreme male brain*. London: Allen Lane.

Baron-Cohen, S. (2007). Sex differences in mind: Keeping science distinct from social policy. In S. Ceci & W. Williams (Eds.), *Why aren't more women in science? Top researchers debate the evidence* (pp. 159–172). Washington, DC: APA.

Bem, S. (1983). Gender schema theory and its implications for child development: Raising gender-aschematic children in a gender-schematic society. *SIGNS: Journal of Women in Culture and Society, 8*, 598–616.

Benenson, J., Tennyson, R., & Wrangham, R. (2011). Male more than female infants imitate propulsive motion. *Cognition, 121*, 262–267.

Berenbaum, S., & Bailey, J. (2003). Effects on gender identity of prenatal androgens and genital appearance: Evidence from girls with congenital adrenal hyperplasia. *The Journal of Clinical Endocrinology and Metabolism, 88*(3), 1102–1106.

Berenbaum, S., & Hines, M. (1992). Early androgens are related to childhood sex-typed toy preferences. *Psychological Science, 3*(3), 203–206.

Berenbaum, S., & Resnick, S. (2007). The seeds of career choices: Prenatal sex hormone effects on psychology sex differences. In S. Ceci & C. Williams (Eds.), *Why aren't more women in science? Top researchers debate the evidence* (pp. 147–157). Washington, DC: American Psychological Association.

Bigler, R., & Liben, L. (2007). Developmental intergroup theory: Explaining and reducing children's social stereotyping and prejudice. *Current Directions in Psychological Science, 16*(3), 162–166.

Bleier, R. (1986). Sex differences research: Science or belief? In R. Bleier (Ed.), *Feminist approaches to science* (pp. 147–164). New York: Pergamon Press.

Breedlove, S., Cooke, B., & Jordan, C. (1999). The orthodox view of brain sexual differentiation. *Brain, Behavior and Evolution, 54*, 8–14.

Burton, F. D. (1972). The integration of biology and behavior in the socialization of *Macaca sylvana* of Gibraltar. In F. E. Poirier (Ed.), *Primate Socialization*. New York: Random House.

Burton, F. D. (1977). Ethology and the development of sex and gender identity in non-human primates. *Acta Biotheoretica, 26*(1), 1–18.

Burton, F. D. (1992). The social group as information unit: Cognitive behaviour, cultural processes. In F. D. Burton (Ed.), *Social processes and mental abilities in non-human primates: Evidences from longitudinal field studies*. Lewiston: Edwin Mellen Press.

Brown, G., Dickins, T., Sear, R., & Lalan, K. (2011). Evolutionary accounts of human behavioural diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*, 313–324.

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review, 106*(4), 676–713.

Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., & Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior & Development, 23*, 113–118.

Douglas, H. (2008). The role of values in expert reasoning. *Public Affairs Quarterly, 22*(1), 1–18.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: Changes in grey matter induced by training. *Nature, 427*(6972), 311–312.

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science, 8*(3), 340–357.

Edelmann, M. N., & Auger, A. P. (2011). Epigenetic impact of simulated maternal grooming on estrogen receptor alpha within the developing amygdala. *Brain, Behavior, and Immunity, 25*(7), 1299–1304.

Escudero, P., Robbins, R. A., & Johnson, S. P. (2013). Sex-related preferences for real and doll faces versus real and toy objects in young infants and adults. *Journal of Experimental Child Psychology, 116*(2), 367–379.

Fausto-Sterling, A. (2005). The bare bones of sex: Part 1–Sex and gender. *SIGNS: Journal of Women in Culture and Society, 30*(2), 1491–1527.

Fausto-Sterling, A., Coll, C., & Lamarre, M. (2012). Sexing the baby: Part 2 – Applying dynamic systems theory to the emergences of sex-related differences in infants and toddlers. *Social Science & Medicine, 74*, 1693–1702.

Fields, R. D. (2010). Change in the brain's white matter. *Science, 330*(6005), 768–769.

Fine, C. (2010). *Delusions of gender: How our minds, society, and neurosexism create difference*. New York: WW Norton.

Fine, C. (2012). Explaining, or sustaining, the status quo? The potentially self-fulfilling effects of 'hardwired' accounts of sex differences. *Neuroethics, 5*(3), 285–294.

Fine, C., Jordan-Young, R. M., Kaiser, A., & Rippon, G. (2013). Plasticity, plasticity, plasticity . . . and the rigid problem of sex. *Trends in Cognitive Sciences, 17*(11), 550–551.

Gettler, L., McDade, T., Feranil, A., & Kuzawa, C. (2011). Longitudinal evidence that fatherhood decreases testosterone in human males. *Proceedings of the National Academy of Sciences, 108*(39), 13194–16199.

Gitau, R., Adams, D., Fisk, N., & Glover, V. (2005). Fetal plasma testosterone correlates positively with cortisol. *Archives of Disease in Childhood. Fetal and Neonatal Edition, 90*, F166–F169.

Golombok, S., & Rust, J. (1993). The pre-school activities inventory: A standardised assessment of gender role in children. *Psychological Assessment, 5*(2), 131–136.

Grossi, G., & Fine, C. (2012). The role of fetal testosterone in the development of "the essential difference" between the sexes: Some essential issues. In R. Bluhm, A. Jacobson, & H. Maibom (Eds.), *Neurofeminism: Issues at the intersection of feminist theory and cognitive neuroscience*. Basingstoke: Palgrave Macmillan.

Haier, R., Siegal, B., MacLachlan, A., Soderling, E., Lottenberg, S., & Buchsbaum, M. (1992). Regional glucose metabolic changes after learning a complex visuospatial/motor task: A positron emission tomographic study. *Brain Research, 570*, 134–143.

Haslam, S., & McGarty, C. (2001). A 100 years of certitude? Social psychology, the experimental method and the management of scientific uncertainty. *British Journal of Social Psychology, 40*, 1–21.

Hassett, J. M., Siebert, E. R., & Wallen, K. (2008). Sex differences in rhesus monkey toy preferences parallel those of children. *Hormones and Behavior, 54*(3), 359–364.

Herman, R. A., Measday, M. A., & Wallen, K. (2003). Sex differences in interest in infants in juvenile rhesus monkeys: Relationship to prenatal androgens. *Hormones and Behavior, 43*, 573–583.

Hines, M. (2010). Sex-related variation in human behavior and the brain. *Cell, 14*(10), 448–456.

Hines, M. (2011). Gender development and the human brain. *Annual Review of Neuroscience, 34*, 69–88.

Hines, M. (2013). There's no good reason to push pink toys on girls. *The Conversation*. Retrieved 10 September 2013 from http://theconversation.com/theres-no-good-reason-to-push-pink-toys-on-girls-15830

Hines, M., & Alexander, G. (2008). Monkeys, girls, boys and toys: A confirmation. Letter regarding "Sex differences in toy preferences: Striking parallels between monkeys and humans". *Hormones and Behavior, 54*, 478–479.

Hines, M., Golombok, S., Rust, J., Johnston, K., Golding, J., & Avon Longitudinal Study of Parents and Children Study Team. (2002). Testosterone during pregnancy and gender role behavior of preschool children: A longitudinal, population study. *Child Development, 73*(6), 1678–1687.

Hines, M., Brook, C., & Conway, G. S. (2004). Androgen and pyschosexual development: Core gender identity, sexual orientation, and recalled childhood gender role behavior in women and men with congenital adrenal hyperplasia (CAH). *Journal of Sex Research, 41*, 75–81.

Hoff Sommers, C. (2012). You can give a boy a doll, but you can’t make him play with it. *The Atlantic*. Retrieved 7 January 2013 from http://www.theatlantic.com/sexes/archive/2012/12/you-can-give-a-boy-a-doll-but-you-cant-make-him-play-with-it/265977/

Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592.

Itani, J. (1959). Paternal care in the wild Japanese monkey, Macaca fuscata fuscata. *Primates, 2*(1), 61-93.

Jadva, V., Hines, M., & Golombok, S. (2010). Infants’ preferences for toys, colors, and shapes: Sex differences and similarities. *Archives of Sexual Behavior, 39*(6), 1261–1273.

Jäncke, L., Gaab, N., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2001). Short-term functional plasticity in the human auditory cortex: An fMRI study. *Cognitive Brain Research, 12*(3), 479–485.

Jordan-Young, R. (2010). *Brain storm: The flaws in the science of sex differences*. Cambridge, MA: Harvard University Press.

Jordan-Young, R. (2012). Hormones, context, and “Brain Gender”: A review of evidence from congenital adrenal hyperplasia. *Social Science & Medicine, 74*(11), 1738–1744.

Kaiser, A. (2012). Re-conceptualizing “sex” and “gender” in the human brain. *Zeitschrift für Psychologie/Journal of Psychology, 220*(2), 130–136.

Kaiser, A., Haller, S., Schmitz, S., & Nitsch, C. (2009). On sex/gender related similarities and differences in fMRI language research. *Brain Research Reviews, 61*(2), 49–59.

Kaplan, G., & Rogers, L. (2003). *Gene worship: Moving beyond the nature/nurture debate over genes, brain, and gender*. New York: Other Press.

Knickmeyer, R., Wheelwright, S., Taylor, K., Ragatt, P., Hackett, G., & Baron-Cohen, S. (2005). Gender-typed play and amniotic testosterone. *Developmental Psychology, 41*(3), 517–528.

Levy, N. (2004). Understanding blindness. *Phenomenology and the Cognitive Sciences, 3*, 315–324.

Li, S.-C. (2003). Biocultural orchestration of developmental plasticity across levels: The interplay of biology and culture in shaping the mind and behavior across the life span. *Psychological Bulletin, 129*(2), 171–194.

Lickliter, R., & Honeycutt, H. (2003). Developmental dynamics: Toward a biologically plausible evolutionary psychology. *Psychological Bulletin, 129*(6), 819–835.

Lombardo, M. V., Ashwin, E., Auyeung, B., Chakrabarti, B., Taylor, K., Hackett, G., et al. (2012). Fetal testosterone influences sexually dimorphic gray matter in the human brain. *Journal of Neuroscience, 32*(2), 674–680.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences of the United States of America, 97*(8), 4398–4403.

Martin, C. L., & Halverson, C. (1981). A schematic processing model of sex typing and stereotyping in children. *Child Development, 52*, 1119–1134.

Mason, W. A. (2002). The natural history of primate behavioral development: An organismic perspective. In D. J. Lewkowicz & R. Lickliter (Eds.), Conceptions of development: Lessons from the laboratory. New York: Psychology Press.

May, A. (2011). Experience-dependent structural plasticity in the adult human brain. *Trends in Cognitive Sciences, 15*, 475–482.

Meyer-Bahlburg, H., Dolezal, C., Zucker, K., Kessler, S., Schober, J., & New, M. (2006). The recalled childhood gender questionnaire-revised: A psychometric analysis in a sample of women with congenital adrenal hyperplasia. *The Journal of Sex Research, 43*(4), 364–367.

Mondragón-Ceballos, R., Chiappa, P., Mayagoitia, L., & Lee, P. (2010). Sex differences in learning the allocation of social grooming in infant stumptailed macaques. *Behaviour, 147*, 1073–1099.

Moore, C. (1984). Maternal contributions to the development of masculine sexual behavior in laboratory rats. *Developmental Psychobiology, 17*(4), 347–356.

Moore, C. (2002). On differences and development. In D. J. Lewkowicz & R. Lickliter (Eds.), *Conceptions of development: Lessons from the laboratory* (pp. 57–76). New York: Psychology Press.

Moore, C., Dou, H., & Juraska, J. (1992). Maternal stimulation affects the number of motor neurons in a sexually dimorphic nucleus of the lumbar spinal cord. *Brain Research, 572*, 52–56.

Muller, M., Marlowe, F., Bugumba, R., & Ellison, P. (2009). Testosterone and paternal care in East African foragers and pastoralists. *Proceedings of the Royal Society B, 276*, 347–354.

Nash, A., & Grossi, G. (2007). Picking Barbie's brain: Inherent sex differences in scientific ability? *Journal of Interdisciplinary Feminist Thought, 2*(1), 1–23.

Nordenström, A., Servin, A., Bohlin, G., Larsson, A., & Wedell, A. (2002). Sex-typed toy play behavior correlates with the degree of prenatal androgen exposure assessed by CYP21 genotype in girls with congenital adrenal hyperplasia. *The Journal of Clinical Endocrinology and Metabolism, 87*(11), 5119–5124.

Orenstein, P. (2011). Should the world of toys be gender-free? *New York Times*. Retrieved 7 January 2013 from http://www.nytimes.com/2011/12/30/opinion/does-stripping-gender-from-toys-really-make-sense.html?_r=0

Pasterski, V., Geffner, M., Brain, C., Hindmarsh, P., Brook, C., & Hines, M. (2005). Prenatal hormones and postnatal socialization by parents as determinants of male-typical toy play in girls with congenital adrenal hyperplasia. *Child Development, 76*(1), 264–278.

Phoenix, C. H., Goy, R. W., Gerall, A. A., & Young, W. C. (1959). Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. *Endocrinology, 65*(3), 369–382.

Reyes, F. I., Winter, J. S. D., & Faiman, C. (1973). Studies on human sexual development. I. Fetal gonadal and adrenal sex steroids. *Journal of Clinical Endocrinology and Metabolism, 37*(1), 74–78.

Rodeck, C. H., Gill, D., Rosenberg, D. A., & Collins, W. P. (1985). Testosterone levels in midtrimester maternal and fetal plasma and amniotic fluid. *Prenatal Diagnosis, 5*, 175–181.

Saad, G. (2012). Sex-specific toy preferences: Learned or innate? *Psychology Today*. Retrieved 7 January 2013 from http://www.psychologytoday.com/blog/homo-consumericus/201212/sex-specific-toy-preferences-learned-or-innate

Servin, A., Bohlin, G., Nordenstrom, A., & Larsson, A. (2003). Prenatal androgens and gender-typed behavior: A study of girls with mild and severe forms of congenital adrenal hyperplasia. *Developmental Psychology, 39*(3), 440–450.

van Anders, S., & Watson, N. (2006). Social neuroendocrinology: Effects of social contexts and behaviors on sex steroids in humans. *Human Nature, 17*(2), 212–237.

van de Beek, C., Thijssen, J. H. H., Cohen-Kettenis, P. T., van Goozen, S. H. M., & Buitelaar, J. K. (2004). Relationships between sex hormones assessed in amniotic fluid, and maternal and umbilical cord serum: What is the best source of information to investigate the effects of fetal hormonal exposure? *Hormones and Behavior, 46*(5), 663–669.

van de Beek, C., Van Goozen, S., Buitelaar, J., & Cohen-Kettenis, P. (2009). Prenatal sex hormones (maternal and amniotic fluid) and gender-related play behavior in 13-month-old infants. *Archives of Sexual Behavior, 38*, 6–15.

van de Waal, E., Renevy, N., Favre, C., & Bshary, R. (2010). Selective attention to philopatric models causes directed social learning in wild vervet monkeys. *Proceedings of the National Academy of Sciences, B, 277*, 2105–2111.

Westermann, G., Mareschal, D., Johnson, M., Sirois, S., Spratling, M., & Thomas, M. (2007). Neuroconstructivism. *Developmental Science, 10*(1), 75–83.

Wong, W., Pasterski, V., Hindmarsh, P., Geffner, M., & Hines, M. (2012). Are there parental socialization effects on the sex-typed behavior of individuals with congenital adrenal hyperplasia? *Archives of Sexual Behavior, 42*(3), 381–391.

Wood, W., & Eagly, A. (2013). Biosocial construction of sex differences and similarities in behavior. *Advances in Experimental Social Psychology, 46*, 55–123.

Neuroethics of Neurodiversity

111

Simon Baron-Cohen

Contents

Abstract

The concept of neurodiversity is relatively recent and has important ethical implications in signaling that there is no single way to be "normal." In this chapter I explore this notion in relation to the neurodevelopmental condition of autism, taking a historical approach to show how attitudes have changed in parallel with changes in who is recognized as having autism and the emergence of an "autism rights" movement.

From Normality to Typicality

If we go back even as far as 20 years ago, we can see that the concept of "normality" was both widespread and categorical. In clinical research, it was common to compare those with a diagnosis to those without, the former being described variously as "patients" or as "impaired" or "pathological," the latter being described as "normal controls." This binary view of the population assumed all those in the "normal control group" were similar to each other (hence categorical) and implied that those in the other group were by definition "abnormal." The clinical and academic profession of

S. Baron-Cohen
Autism Research Centre, Psychiatry Department, Cambridge University, Cambridge, UK
e-mail: sb205@cam.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_150,

psychology even had a whole subdiscipline called “abnormal psychology” with textbooks dedicated to this view of those who were not normal.

What changed was both subtle and gradual. In terms of subtlety, the term “normal” gradually gave way to the less value-laden term “typical” as a way of describing those without a diagnosis. It is a more neutral term because it describes some statistical norm or average, which by definition is more common than those at the extremes. In this sense, “typical” simply means more common, and the contrast group is no longer referred to as “abnormal” but as “atypical.” (In recognition of this, in 2008 I changed the title of the course I teach at Cambridge University from its historical name “abnormal psychology” to its current name “atypical psychology.”). A second way in which concepts subtly changed was to connect with the field of *individual differences*, previously mainly restricted to those who studied the dimension of IQ or those who studied personality dimensions. The importance of this was to acknowledge differences even within the typical group and frequently to recognize that those who have a diagnosis are simply an extreme on a dimension of individual differences. This marked a moving away from categorical diagnosis to dimensional diagnosis.

The Autism Spectrum

A very clear example of this is in the field of autism. Autism is a neurodevelopmental condition of partly genetic origin, affecting brain growth, structure, and function. In turn it leads to altered information processing, causing a disability in social relationships and communication but a remarkable attention to detail, unusually narrow interests, a strong need for repetition and routine, and an aptitude for detecting patterns.

Previously it was easier to maintain the belief that autism was categorical because most cases identified not only showed the above features of autism but they also had below average IQ (therefore learning difficulties) and language delay. With this “triple hit” of disability, it seemed plausible that they might be qualitatively distinct from the rest of the population. But during the 1990s, this categorical view became harder to sustain, not least because individuals with autism were being identified at every subtle gradation on the IQ range. While the convenience was to define someone with autism as having delayed language (not speaking by 2 years old) and frequently below average IQ (more than 2 standard deviations below the population mean of 100, so below 70), it became more and more difficult to defend this portrait of autism. This is because there were individuals who had an IQ above this level and who were (for a time) referred to as “high functioning.” In parallel with “admitting” more and more higher functioning individuals into the category of autism, we were also “admitting” individuals with no trace of language delay (they might even have had the opposite pattern, of speaking precociously, with an adult-like vocabulary) and even with an above-average IQ.

“Asperger syndrome” was the name given to individuals with the core signs of autism but with average IQ (or above) and intact language development.

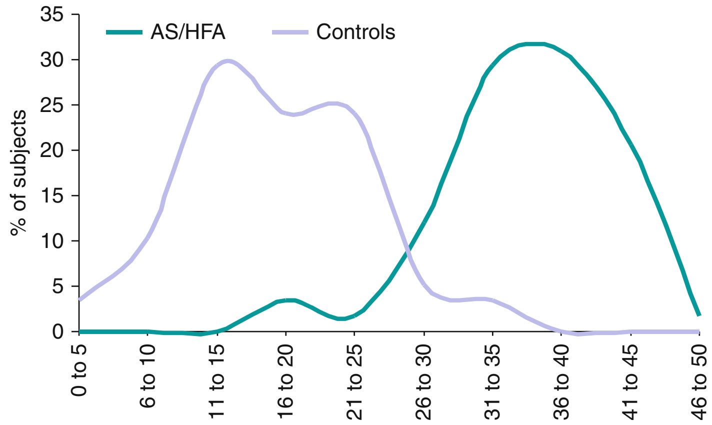

Fig. 111.1 The Autism Spectrum Quotient (AQ)

Uta Frith's book entitled "Asperger Syndrome" and published in 1991 was perhaps the first book in English on this subject and included a translation of Hans Asperger's description of the children who today have a diagnosis of Asperger syndrome. This followed on the wave of interest generated by Lorna Wing's (1981) article in Psychological Medicine introducing Asperger syndrome to the English-speaking biomedical community. But the gradual effect of clinics worldwide starting to diagnose children and adults with Asperger syndrome, as part of the autism spectrum, was to drive home the individual differences approach (such individuals seeming to be the bridge between classic autism and those without a diagnosis) as well as to raise the serious possibility that Asperger syndrome was nothing more than the extreme end of a single dimension.

The publication of the Autism Spectrum Quotient (AQ) in 2001 was one of the first clear demonstrations that autistic traits lie on a continuum on which we can all be scored (Baron-Cohen et al. 2001). The AQ, a questionnaire which probes five different domains of behaviors, preferences, and interests including social skills, attention switching, communication, imagination, and attention to detail, is normally distributed, and – as shown in Fig. 111.1 – those without a diagnosis lie under the left-hand curve, while those with a diagnosis lie under the right-hand curve. Most importantly, these two curves overlap, reminding us that there is no clear point at which those with and without a diagnosis can be neatly separated. This offered strong confirmation of what Lorna Wing had in 1988 called "the autistic spectrum."

Different Ways to Be "Normal"

If in the population people simply vary in terms of the number of autistic traits they have, in what sense can we call someone "normal" and someone else "abnormal"? The very notion collapses as people without a diagnosis differ from each other as much as people with a diagnosis. And of course, autistic traits are not the only kind

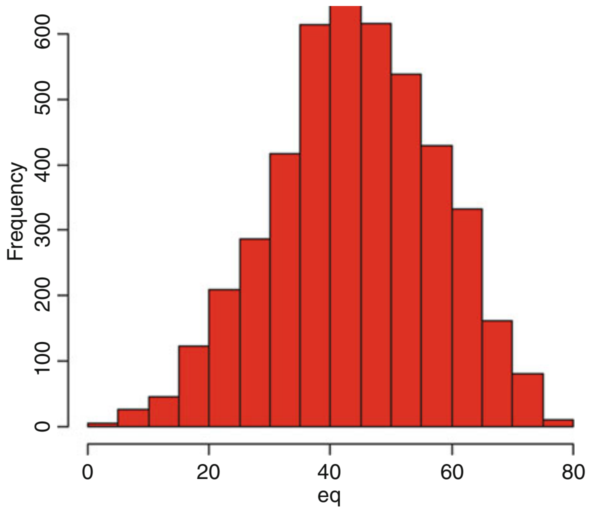

Fig. 111.2 Data from the Empathy Quotient (EQ)

of "diversity" that one finds in the population. The first example includes handedness (left handedness is thankfully no longer seen as "abnormal" or "sinister," and children are thankfully no longer forced to write with their right hand, as I was as a child), and even the notion of left vs. right handedness is challenged, as it has become clearer that being strongly right handed and left handed are just two possible positions on a dimension of individual differences, ambidextrous individuals being a third position, but with people in the population being found in intermediate positions between these clearer points. The second example is sexual orientation, where "gay" or "straight" turn out to be too simplistic a binary classification, given the existence of "intermediate" forms (bisexuality, other). The third example is how easily a person can empathize (reflected in the empathy bell curve that results from using another measure, the Empathy Quotient (EQ), in the population (see Fig. 111.2). And the final example would be individual differences in systemizing drive (the drive to analyze or build a system), as reflected in data from the Systemizing Quotient (SQ).

Given that all of these dimensions of individual differences correlate with differences in brain structure and function (e.g., Lai et al. 2011), it must mean that there are many different ways to develop, many different routes to adulthood, and many ways to function, hence the notion of neurodiversity.

Extending Neurodiversity to Those with a Clinical Diagnosis

If typical individuals manifest neurodiversity, it is only a small step to extend this notion to those with a diagnosis. The radical but quite reasonable proposal from the neurodiversity movement is that those with a diagnosis are simply developing differently, not necessarily worse (or better) than others, but simply different.

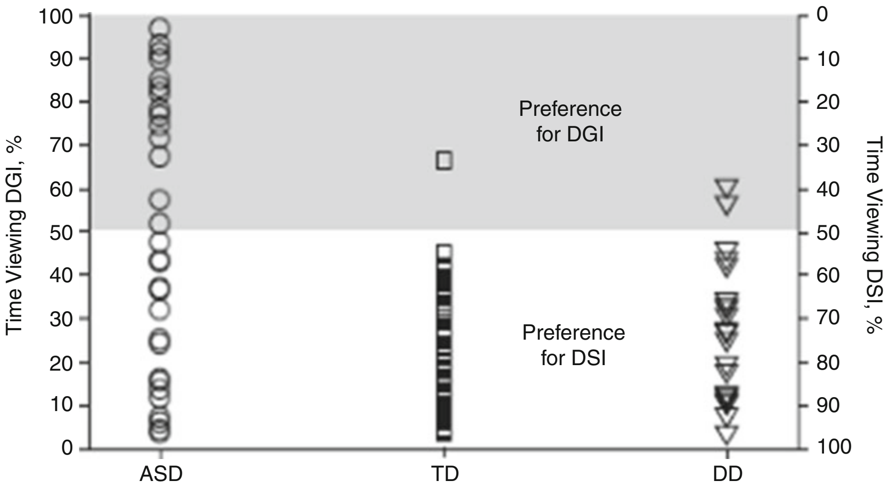

Fig. 111.3 From Pierce et al. (2011)

The benefit of this view is that it makes space for precocious rates of development just as much as delayed rates of development, and it avoids value judgments implicit in the terms "normal" and "abnormal." If we continue to consider autism as our example, we can view children or adults who receive this diagnosis as simply different in their patterns of interests.

A nice study from Karen Pierce and colleagues in San Diego measured how long toddlers with and without autism looked at social (e.g., pictures of children playing) vs. a nonsocial stimuli (e.g., moving geometric figures). They found that more toddlers with autism ("ASD" in the graph) looked for longer at the nonsocial stimulus, and more typically developing toddlers ("TD") looked for longer at the social stimulus. Toddlers with developmental delay ("DD") also showed the typical pattern of a social preference. This study is a clear example of how children with autism are simply showing a difference in their patterns of interest. Looking at a social stimulus may be the typical pattern of attention, but not conforming to this pattern is not necessarily a sign of being worse or impaired, since it is simply reflecting that toddlers with autism allocate their attention to different aspects of the environment. See Fig. 111.3.

Neuroethics of Neurodiversity

So what are the benefits of thinking in terms of neurodiversity? The most obvious benefit is in terms of inclusion and acceptance. Instead of those with a diagnosis feeling in some way inferior or excluded, the notion of neurodiversity breaks down divisions between those with and without a diagnosis by acknowledging that people

are simply "differently wired" and that it is misplaced to think of some people's wiring as "normal." When I was a child, "left handers" were still regarded as abnormal and coerced to write with their right hand as the only right way (no pun intended). I am an example of this kind of educational antidiversity. A second ethical implication is in terms of human rights, where those with autism, for example, have applied the same human rights framework to those with a disability as might be applied to any other minority group (e.g., based on ethnicity). The positive effect of this has been for the majority "culture" to think of how the environment can be adapted to avoid any risk of discrimination. Just as traffic lights need to be designed to be heard as well as seen, to improve their access to the blind as well as deaf pedestrians, we can imaginatively consider all parts of our environment, both social and physical, to make it equally accessible to a diverse range of neurologies. A third ethical implication is that because someone is neurologically different, this does not mean they need a treatment or a cure. They may need support (people with autism are a good case in point who need support for their social disability), but their difference includes their excellent attention to detail (Jolliffe and Baron-Cohen 1997) and treating them for their autism might risk them losing such superior skills.

However, we have to consider the limits of the notion of neurodiversity. One is that not all kinds of neural wiring are equally good, since some actually cause "disease." Examples would be Parkinson's disorder, epilepsy, or dementia. These cause unequivocal impairment, where the ethical option is treatment if this is available, to improve the quality of a person's life. Uncontroversial "targets" for treatment in some people with autism include certain "comorbid" symptoms that are sometimes present in autism such as epilepsy, gastrointestinal pain, self-injury, and even severe learning difficulties. Even aptitudes such as social skills and communication are reasonable targets for treatment and intervention. As with all interventions, the priority has to be ensuring the benefits are not outweighed by any unwanted side effects. In the case of autism, the goal should be to identify interventions that target the areas of clear disability, whilst leaving the areas of difference to blossom and reach their full potential. How to assess functioning and flourishing in people who are neurologically different and where to draw the line between disability and difference are questions that professionals need to continuously revisit (Fenton and Krahn 2007; Jaarsma and Welin 2011). We need to embrace the notion of neurodiversity, but without applying it unthinkingly to all forms of neurological difference.

References

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17. doi:10.1023/A:1005653411471. PMID 11439754.

Fenton, A., & Krahn, T. (2007). Autism, neurodiversity and equality beyond the 'normal'. *Journal of Ethics in Mental Health, 2*(2), 1–6.

Jaarsma, P., & Welin, S. (2011). Autism as a natural human variation: Reflections on the claims of the neurodiversity movement. *Health Care Analysis, 20*(1), 20–30.

Jolliffe, T., & Baron-Cohen, S. (1997). Are people with autism and asperger syndrome faster than normal on the embedded figures test? *Journal of Child Psychology and Psychiatry, 38*(5), 527–34.

Pierce, K., Conant, D., Hazin, R., Stoner, R., & Desmond, J. (2011). Preference for geometric patterns early in life as a risk factor for autism. *Archives of General Psychiatry, 68*(1), 101–109. doi:10.1001/archgenpsychiatry.2010.113.

Wing, L. (1988). The continuum of autistic characteristics. In E. Schopler & G. Mesibov (Eds.), *Diagnosis and assessment in autism*. New York: Plenum Press.

Section XXIII

Weaponization of Neuroscience

Weaponization of Neuroscience

112

Gerald Walther

Contents

Abstract

This introductory chapter to the section on the "Weaponization of Neuroscience" provides a short overview of the developments in biological disarmament since the beginning of the twentieth century. It starts with an account of the problems at the Fifth Review Conference of the Biological and Toxin Weapons Convention to establish a verification regime and how the Convention subsequently shifted its focus from state to non-state actors in terms of preventing the development of biological weapons. As part of this reorientation, the issue of the potential misuse of benign scientific research was raised, which was coined "dual-use biosecurity" or the "dual-use dilemma" in the following debates. While the international security community was initially only worried about the misuse of neuroscience by malign non-state actors, the disarmament community has expanded the topic to include the use of scientific research by state military as well.

The five chapters in this section discuss the following: Is neuroscience research of interest to the military? Does international law regulate neuroscience? What is the level of awareness and education of neuroscientists on the topic of dual use? How can the neuroscience community work towards protecting their research from misuse? What can neuroethics do?

G. Walther
Division of Peace Studies, University of Bradford, Bradford, UK
e-mail: g.walther@bradford.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_144,

Biological Disarmament in the Twentieth Century and the Origins of Dual-Use Bioethics

It is probably not farfetched to argue that the events of 9/11 had a profound impact on international politics. However, in the shadows of 9/11, the anthrax attacks that followed proved to be equally important to the security community and particularly to those interested in biological disarmament. Just prior to 9/11, the Biological and Toxin Weapons Convention (BTWC) suffered a significant setback in strengthening its ability to ensure a biological weapon-free world. The BTWC, albeit already established in 1972, lacks a regime that oversees whether signatory states actually comply with the regulations set out in the Convention, which is primarily to prohibit any work that could be used for the production, stockpiling, or delivery of biological weapons. In 2001, at the fifth review conference, a proposal had been put forward by the then Chair Tibor Tóth to remedy this lack. However, the USA eventually refused their support for this proposal and questioned the utility of trying to come up with a verification regime altogether. In this political disaster, 9/11 and the anthrax letters fell. As part of a way forward for the BTWC, the topic was opened up of how likely biological research could be misused for malign purposes, i.e., bioterrorism, and what could be done to decrease the prospect of it. The label for this discussion was "dual-use biosecurity."

The discussion was heavily influenced by the work of the so-called Fink Committee, or in full the "Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology," which worked between April 2002 and January 2003 to produce a report published in 2004 by the US National Academy of Science. The Fink Committee identified two issues. First, domestic and international legislation ensures the physical protection of laboratory workers and the environment from harmful or novel biological agents. However, these regulations do not address the issue of how research could be protected from being abused and misused by malign actors. There is also no method or process on how to weigh the potential dangers of misuse against the beneficial aspects of research. The Fink Committee proposed to address this issue by engaging the scientific community in a dialogue as well as to establish oversight committees that evaluate scientific work on a danger vs. benefit scale. In the committee's view, scientists have a "moral duty to avoid contributing to the advancement of biowarfare and bioterrorism" (NRC 2004, p. 112). While this debate on the role and responsibility of scientists was going on within the international security community, i.e., primarily within the BTWC, both scientists and bioethicists were largely silent on the issue. Of course there were some initial reactions, for example, several editors of major biology journal, who had been consulted by or had been members of the Fink Committee, issued a joint statement in 2003, where they promised to take into consideration biosecurity and biodefense concerns when publishing a paper. Nevertheless, as Minehata and Walther describe in ▶ Chap. 113, "Biosecurity Education and Awareness in Neuroscience" of this section, this discussion of the duty of scientists did not extend or involve the majority of the scientific community. Equally, it took until 2009 for Dando to note that the

bioethicists had finally entered the dual-use debate (Dando 2009). This claim was based on three articles (Kuhlau et al. 2008; Ehni 2008; Miller and Selgelid 2008) that discussed the moral responsibility of scientists and what they could be expected to do in order to reduce security concerns about their work. Except for these three articles though, dual-use bioethics has not attracted much further attention from bioethicists. The debate is thus largely dominated by the security community and has yet to reach the widespread attention of practicing scientists or bioethicists. However, in 2011, the case of the H5N1 research has shown how contentious the issue may become as the security community is trying to change legislation and increase both oversight and regulation of scientific work. In the H5N1 case, a mammalian-transmissible H5N1 influenza was created, which raised the question of whether this research could have been considered as offensive biological weapons research and if the risks associated with publishing the research outweigh any potential benefits (Novossiolova et al. 2012). Opinions on this issue were quite divided between the security and science community, and even internally there was no consistency, so this topic might spark further debates in the future.

Section Overview

The first chapter in this section by Minehata and Walther highlights the current awareness of the dual-use topic among scientists and particularly neuroscientists. They advocate the use of education to help the science community deal with the security implications of their work, which includes the use of research by the military, before it becomes a political issue (Minehata and Walther). After all, it cannot be in the interest of the scientific community to have unaided lawmakers start to regulate science. However, as the chapter points out, current awareness of this issue is very limited among scientists, and it will presumably take time for this issue to become a topic of debate. Education programs that address this issue are very limited at the moment. Only the University of Bradford in the UK has a training module that specifically teaches the issue of dual-use bioethics to practicing scientists (http://www.brad.ac.uk/bioethics/). While the project has been running successfully since 2010, the number of its graduates is just over 100 so the impact is as yet limited. In the neurosciences, a series of workshops have been held at the University of Manchester and the University of Bradford in 2012 and 2013 to develop an ethics course for neuroscientists, which will include a component on dual-use bioethics as well (http://www.lab.ls.manchester.ac.uk/neuroethicseducation/). Within neuroethics, the issue of the relationship between the military, national security, and neuroscience has been discussed in a target article and several open peer commentaries in Volume 1, Issue 2, of the American Journal of Bioethics Neuroscience (Marks 2010) and in a booklet by The Royal Society (2012). It has also featured as a panel on “Neuroscience, National Security, and Society” at the 2011 meeting of the International Neuroethics Society meeting.

As discussed by panelist James Giordano, one of the problems with discussing the potential malign uses or misapplications of neuroscience is a lack of pragmatic

consideration of what is actually feasible (http://www.neuroethicssociety.org/2011-meeting-archive). Popular science fiction, for example, the idea of a "Manchurian Candidate," may cloud what can actually be accomplished and influence public opinion and understanding in excess of the facts. In ▶ Chap. 114, "Neuroscience Advances and Future Warfare" in this section, Dando takes a look at history of neuroscience and the military with regard to the development of lethal nerve gases. From there, he discusses the potential of misuse or military use of contemporary neuroscience developments and neurotechnologies. The primary argument is that the emphasis on misuse, e.g., bioterrorism, is misguided and that in the sort of asymmetric warfare in the coming decades, a militarization of the life sciences is more likely and needs to be monitored. As examples of current technology, Dando looks at transcranial magnetic stimulation and brain-computer interfaces and their implications for information processing and autonomic weapons systems and how these will pose challenges to our understanding of morality and international law.

The question of the interaction between international law and the militarization of neuroscience is taken up specifically by Jefferson in ▶ Chap. 115, "International Legal Restraints on Chemical and Biological Weapons" of this section. She discusses international conventions such as the BTWC and the Chemical Weapons Convention with a specific focus on how they regulate incapacitants, which act on the nervous systems and therefore could be considered neuroweapons. Even though Jefferson concludes that international conventions cover neuroweapons such as incapacitants, she agrees with Dando that education of neuroscientists is necessary to raise awareness in order to reduce the likelihood that neuroscience will be militarized.

However, as has been shown in the Minehata and Walther chapter, despite efforts to promote ethics education from State Parties, there has been little progress made thus far. Novossiolova takes up this challenge and discusses how biosecurity needs to be understood in the context of changing norms and values within an epistemic community.

The final chapter of the section shifts the focus by analyzing how neuroscientific insights could have an impact on international law. Walther challenges current international humanitarian law and its focus on preventing physical harm to civilians and improving the security of combatants. Based on a review of how fear, anxiety, and stress impact neurophysiology and can cause long-term or permanent damage to the individual, he argues that recent military strategies such as "shock and awe," as used in the 2003 Iraqi invasion, as well as military interest in "fear creation" may lead to more psychophysical harm. He concludes that neuroethicists should use their understanding of neuroscience to inform and change humanitarian law in the future.

Conclusion

The militarization of neuroscience is not a novel issue. The twentieth century saw several neuroweapons in deployment, starting with the development and use of lethal nerve gases in the World Wars. However, with the surge in our understanding

of the brain and the nervous system since the late twentieth century, the potential for use by malign actors as well as the military has increased considerably. And it is not only the fear of science-fiction dystopias that the neuroethics and neuroscience community should actually be concerned about. Current neurotechnologies already offer possibilities of misuse, and these are more readily accessible than ever before. The public needs to be assured that neuroscience does not open a Pandora's box but that the implications for our open civil society are not threatened by these developments. This section will argue that education and awareness-raising are two tools to allow neuroscientists to understand their impact on society as well as enable them to evaluate the risks of their research. These questions might be best served by a conscious deliberation by individual scientists rather than a sole reliance on Ethics Review Boards.

Cross-References

► Biosecurity as a Normative Challenge
► Biosecurity Education and Awareness in Neuroscience
► International Legal Restraints on Chemical and Biological Weapons
► Neuroethics of Warfare
► Neuroscience Advances and Future Warfare

References

Dando, M. R. (2009). Bioethicists enter the dual-use debate. Bulletin of the atomic scientists. http://www.thebulletin.org/web-edition/columnists/malcolm-dando/bioethicists-enter-the-dual-use-debate. Accessed 29 Jan 2013.

Ehni, H.-J. (2008). Dual use and the ethical responsibility of scientists. *Archivum Immunologiae et Therapiae Experimentalis, 56*, 147–152.

Kuhlau, F., Eriksson, S., Evers, K., & Höglund, A. T. (2008). Taking due care: Moral obligations in dual use research. *Bioethics, 22*(9), 477–487.

Marks, J. H. (2010). A neurosceptic's guide to neuroethics and national security. *AJOB – Neuroscience, 1*(2), 4–12.

Miller, S., & Selgelid, M. J. (2008). *Ethical and philosophical consideration of the dual-use dilemma in the biological sciences*. New York: Springer.

National Research Council. (2004). *Biotechnology research in an age of terrorism*. Washington, DC: National Academies Press.

Novossiolova, T., Minehata, M., & Dando, M. R. (2012). The creation of a contagious H5N1 Influenza virus: Implications for the education of life scientists. *Journal of Terrorism Research*, *3*(1), Special Issue – Assessing the Emergency Response to Terrorism. http://ojs.st-andrews.ac.uk/index.php/jtr/issue/view/46. Accessed 29 Jan 2013.

The Royal Society. (2012). *Brain waves module 3: Neuroscience, conflict and security*. London: The Royal Society. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/brain-waves/2012-02-06-BW3.pdf. Accessed 29 Jan 2012.

Biosecurity Education and Awareness in Neuroscience 113

Masamichi Minehata and Gerald Walther

Contents

Abstract

This first content chapter of the section starts with a short historical overview of the military interest in neuroscience research. Based on this account, it asks the question of how neuroscience has responded to the ethical issues that have emerged as part of these research activities. An analysis of current ethical training shows that such a debate has been underdeveloped and that the question of dual-use within ethics education of neuroscientists is very limited. Based on these findings the chapter explores the relationship between the security community, specifically, the Biological and Toxin Weapons Convention, and how neuroethicists could help the Convention in addressing the lack of awareness within the neuroscience community of the militarization and dual-use of their research and technologies.

M. Minehata (✉) • G. Walther
Division of Peace Studies, University of Bradford, Bradford, UK
e-mail: m.minehata@hotmail.com; G.Walther@bradford.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_141,

Introduction: The Historical Dimension of Military Interest in Neuroscience

While the military community has become very interested in recent advances in neuroscience research, this awareness of the potential of neuroscience for warfare is far from novel. Already in the early stages of modern neuroscience during and after World War II, research into functional neuroanatomy was of keen interest to the military. For example, Jose Delgado's work in the 1950s and 1960s, which was funded by two agencies within the Department of Defense (Snyder 2009), including the Office of Naval Research (Horgan 2005), attracted attention because of his experiment in 1964 in Spain where he was able to stop the charge of a "black bull," a breed of bulls used for bullfighting, toward him by flipping a switch that activated an electrode implanted in the brain of the bull. While it is inconclusive as to whether Delgado was able to reduce the aggression or simply inhibited the movement of the bull in this specific experiment, which nevertheless received considerable attention by the media, the *Time Magazine* has already reported in 1953 that Delgado had been able to reduce aggression in macaque monkeys (Snyder 2009).

In addition to this interest in the open scientific community, the military's own, top secret program "MK-ULTRA" also used neuroscientific and psychological insights and techniques for a variety of purposes. Unfortunately, nearly all documents of its existence have been erased. The program was established in the early 1950s and continued until 1973 and used a variety of tools, such as the use of psychopharmacological agents, e.g. LSD, or psychological methods, e.g. hypnosis, to alter the state of minds of soldiers for beneficial, i.e. producing "supersoldiers," or malign purposes, i.e. controlling behavior or improving torture techniques (Streatfeild 2006). In 1965, seven psychologists delivered a classified paper to the Pentagon titled "Phenomena Applicable to the Development of Psychological Weapons," which outlined "how arms makers could enhance the psychological effect of new weapons, how fear could be promoted, how perceptions could be altered, how the stress of combat could be increased, and how the psychological differences among racial groups could be exploited" (Moreno 2012, p. 98).

Besides this interest in functional neuroscience, pharmacological agents to kill or demobilize the enemy have of course already been used in World War II, where Germany used the nerve gas VX. While, as discussed by Jefferson in ▶ Chap. 115, "International Legal Restraints on Chemical and Biological Weapons," the *Biological and Toxin Weapons Convention* (*BTWC*) prohibits the use of biological weapons, some agents acting on the nervous system may only be covered by the *Chemical Weapons Convention* (*CWC*), which allows the use of riot control agents, which mainly affect the nervous system, for domestic law enforcement purposes. In 2002, the Russian police used a nerve gas to incapacitate terrorists that held several hundred people as hostages in a Moscow theater. The outcome of the use of the incapacitant was 117 deaths as well as permanent damage to others (Schliermeier 2002).

Given this long history in the work of neuroscientists, it may be presumed that the community has developed a code of conduct for their research behavior or at least discussed the implications of their research for society and the military. The next section will analyze if this hypothesis can be validated.

Ethical Training in Neuroscience is Underdeveloped

In a commentary in *Science* in 2009, Barbara Sahakian and Sharon Morein-Zamir urged the neuroscientific community to introduce ethical training to neuroscience curricula because advances in neuroscience will have important ramifications for "*education*, treatment, and the law" (Sahakian and Morein-Zamir 2009). In their informal survey among 20 major research-intensive universities in the UK, they found that little formal ethical training is given to students (Sahakian and Morein-Zamir 2009). In Canada on the other hand, a telephone survey conducted in 2010 provided an opposite picture (Lombera et al. 2010). In order to verify if there is a possible discrepancy between the educational systems, Walther (2012) conducted a survey in several countries – Australia, Canada, Germany, the UK, and the USA – of both undergraduate and postgraduate neuroscience courses.

The survey first looked at information available at department web pages about ethical training, which was then supplemented by a questionnaire sent to the course administrators. Combining both types of data showed that 52 % of all courses had some form, i.e. formal or informal, of ethics training compared to 80 % when just taking into consideration the questionnaire responses. Formal training was given in 63 % and 60 %, respectively, of the courses that had ethical training. In terms of a possible variation between countries in terms of ethical training, the survey concluded that all countries had similar results. The only exception was the UK, specifically with regard to the question if formal training is given in the courses. All ethical training provided in the UK was delivered in informal sessions. Walther argues that a form of "mainstreaming" has taken place in the UK, which is a term used within business ethics education, where ethical issues are discussed in an ad hoc manner rather than having an exclusive course on the issue. The idea behind it is to bridge the gap between training and practice. Of course, explaining that ethical training has been embedded within all of the modules of a course may simply be a more sophisticated way of saying that ethics is actually not being taught.

When asked about the issues that are addressed in the ethical training, Walther reports that the most common issues were "human and animal use in research," "research misconduct," "treatment of data," "authorship and allocation of credit," and "the researcher in society." While the first four may be considered standard issues within the ethics of science, the last entry of the role of the researcher in society might constitute a discussion of the implications of research on society, which would cover issues of militarization of research and the potential for it to be misused by malign actors. However, the topic rather stresses the role that a scientist has as an expert in a debate and how scientists can reconcile their professional duty to be objective with their subjective wishes, convictions, and beliefs. Indeed, in the

survey the phrase dual-use was only mentioned in 6 % of all of the courses in the sample.

Indeed, proposed solutions to the topic of misuse of research by malign actors as discussed in the *BTWC* arena, where the focus is on malign actors such as terrorists and not misuse by the military of States Parties, may run counter to prevailing practices of scientific publishing. For example, in order to reduce the likelihood that research can be misused by these malign non-state actors, it has been suggested to reduce the amount of information available within publications in order to make it more difficult to replicate the experiments and, for example, be able to replicate a particular pathogen or modify existing ones to make them more virulent. However, the *Journal of Neuroscience* published by the Society for Neuroscience clearly states in its "guidelines for authors of scientific communications" that they are required to make all "unique and propagatable" material accessible to anyone who asks for them under article 1.12 (Society for Neuroscience 2012). The only reservations are where commercial interests are at stake. These rules are part of the scientific canon of the freedom of scientific inquiry and the collegial nature of a communal inquiry for scientific explanations and the generation of knowledge. Thus, any interference with the freedom to publish will be seen as an attack on the values of the community that have yielded remarkable success in modern times.

Having considered the lack of awareness, *education*, and policy development to promote awareness raising of neuroscientists about dual-use issues, it might prove useful to inquire about the state of education in the wider field of the life science. The following sections expand an analytical scope to the current state of education of *dual-use* issues in the life sciences.

Necessity for Awareness Raising of Life Scientists

A major challenge for today's international community has been to find effective ways to raise awareness among life scientists about their social responsibility regarding the potential for the destructive use of the life science research in which they are engaged. The life sciences differ from nuclear science developments in that they are conducted around the world in commercial and academic laboratories rather than those belonging to national governments (National Research Council 2006). In addition to this wider scale of practice, the actual speed of scientific advancement and resulting security implications are "possibly too fast for any State, organization or individual to cover alone" (Millet 2010, p. 29). Moreover, there are critical ambiguities surrounding the boundary between defensive and offensive biological programs which can be used to blur issues of legality (under the *BTWC* that the development of all biological weapons is illegal, as is their production, acquisition, transfer, retention, stockpiling, and use). Finally, in order to address the concerns of scientists, approaches aiming to promote a culture of biosecurity-based social responsibility need to be mindful to "ensure a focus on the highest-risk research and avoid unnecessary restrictions or censorship" over scientific freedom (Smith et al. 2010, p. 137).

Accordingly, there is a need for better collaboration between scientific communities and policy-makers. For this very reason, there is also a need for *education* specifically designed to better inform scientists' and policy-makers' understanding of how the potential for the misuse of the life sciences and related technologies can be recognized and prevented. In this chapter, such efforts to develop a web of preventive policies are broadly envisaged as *biosecurity* (Feaks et al. 2007). It is suggested that such biosecurity education should incorporate themes such as, inter alia, the potential for dual-use risks in contemporary life sciences; the responsible conduct of research and ethical approaches among life scientists; the history of biological warfare programs and biological terrorism; the role of international prohibition regimes and their national implementation (such as the BTWC); the intersection of public health and national security; and the building of an effective set of preventative policies to ensure the security of benign developments in the life sciences.

Increased International Attention and Remaining Problems

The necessity of awareness raising among life scientists about dual-use issues has been underlined by governments and professional communities in science, public health, security, and ethics (Miller and Selgelid 2007; Gorski and Spier 2010). These include the InterAcademy Panel (IAP) (Inter-Academy Panel 2005; Nature 2003), the World Health Organization [WHO] (2007), the Organization for Economic Cooperation and Development [OECD] (2007), and the BTWC (2008). However, against such growing international attention towards *biosecurity education*, the lack of awareness of individual scientists across the globe has been clearly demonstrated (Dando and Rappert 2005; Rappert et al. 2006). Further to this, the limited availability of biosecurity tutoring at the university level has been illustrated by a series of international surveys investigating the current state of biosecurity-related degree courses in the United States (National Research Council and American Association for the Advancement of Science 2009), Europe (Mancini and Revill 2008), Japan (Minehata and Shinomiya 2010), Israel (Freedman 2010), and the Asia-Pacific region (Minehata 2010). In the case of Japan, for example, it was noted that a range of difficulties were faced by university lecturers in introducing appropriate teaching due to:

- An absence of space in existing curricula
- An absence of time and resources available to develop new curricula
- An absence of expertise and available literature on biosecurity education
- General doubt and scepticism about the need for biosecurity education on the part of educators and scientists (Minehata and Shinomiya 2010)

Despite these commonly experienced obstacles, the surveys indicated the possibility of promoting biosecurity education by utilizing already-implemented ethics education processes. We found that a large number of universities surveyed already have educational modules focusing on ethics. Therefore, it has been suggested an expansion of the scope of traditional ethics education by integrating the concept

of dual-use biosecurity as part of the *education* in social responsibility already offered to life scientists (National Research Council and American Association for the Advancement of Science 2009; Revill et al. 2009).

What Is Needed?

A good strategy is needed to deal with the range of obstacles by showing that biosecurity education may be implemented and that it need not be expensive, time-consuming, or onerous in terms of workload. What kind of educational material can be of use in the absence of widespread expertise and available literature on *biosecurity education*? One answer is to place open-source teaching materials online, via the Internet. There are a number of benefits to this approach. Firstly, it is important to recognize that there is no "one-size-fits-all" answer to biosecurity education. Secondly, there are significant differences in priorities between academic institutions, national and regional needs, and also variations derived from differing sociocultural backgrounds. Because of these distinctions, online educational resources are useful, as they can be modified and tailored by users in order to fit the specific teaching modes and needs in various local educational contexts. Further, they ease pressures on time spent in planning and preparing material, overcome financial constraints on the development of biosecurity curricula, and provide the expertise required for efficient and effective assimilation of such material (Dando 2008). In addition, tutors are free to choose what to include in sessions and at what level to "pitch" their teaching, depending on the educational level and technological perspectives of the audience.

Following on from the use of this type of educational resource, the next stage in promoting biosecurity education is to build education capacity through the implementation of similar modules in different academic contexts and institutions. This will disseminate knowledge and understanding more broadly and is likely to engage a wider range of students, educators, and scientists as it is implemented.

Who Should Be Involved?

At the Meeting of Experts of the *BTWC* in 2008 during the discussion on education, States Parties addressed this view by noting that target audience of dual-use education should include medical institutions, scientific organizations, professional institutions, companies, policy-makers' administrators, and universities and authors, editors, and publishers (Japan in consultation with JACKSNNZ 2008, p. 2; Netherlands 2008, p. 6). Indeed, the significance of raising awareness about the ethical responsibility of scientists in preventing the misuse of the life sciences, and to preserve scientific autonomy, has been widely discussed in literature (Atlas and Dando 2006; Kuhlau et al. 2008; Selgelid 2007; Somerville and Atlas 2005). It is

necessary that scientists assume their social responsibility for any dual-use consequence of their research. This is because scientists are at the forefront of scientific understanding and are thus best placed to envisage the consequences of their research. This indicates their primary responsibility in extending the culture to "do no harm" into the area of dual-use against humans, animals, and plants. They are the primary actors who can "help ensure a focus on the highest risk research and avoid unnecessary restrictions or censorship" (Smith et al. 2010, p. 137).

At the same time, nonscientists have the ethical responsibility of doing no harm against the freedom of scientific research. Scientists are creating innovative knowledge with a view to making social benefits in terms of medicine, the environment, and socioeconomic benefits which other social actors can enjoy. The freedom of scientific inquiry, as well as the freedom to share knowledge including the form of publications, is a fundamental human right. Even if a possibility for the misuse of science exists, security regulation over peacefully designed scientific research requires careful dialogue between scientists and nonscientists.

Awareness about dual-use issues is not only the matter of ethics of scientists, but it is more about the practical interest of professional development, such as publication or funding of scientific research. For example, the combination of voluntary awareness raising of scientists about dual-use issues by scientific community, national legislation on *biosecurity* and biosafety measures, and oversight of scientific research was underlined by the "Statement on the Consideration of Biodefence and Biosecurity" signed by the editors of some 30 privileged scientific journals, including *Science*, *Nature*, and *Cell*, in 2003 (Nature 2003) as well as by the "Statement on Biosecurity" of the InterAcademy Panel which was endorsed by national science academies of more than 60 states in 2005 (IAP 2005). In 2008, the US "National Science Advisory Board for Biosecurity" (NSABB) introduced a strategic plan "to provide recommendations on developing programs of outreach and education on dual-use research issues for all scientists and laboratory workers at federally-funded institutions."

Fukushi (2012) illustrated that there are potential intervention points to promote awareness raising of neuroscientists about *dual-use* issues. Regarding the role of journal editors, "*Neuroethics*," "*Journal of Cognitive Neuroscience: Neuroethics Section*," and the "*American Journal of Bioethics: Neuroscience*" can be great places to address issues alongside the journals which recognize the issue with the "Statement on the Consideration of Biodefence and Biosecurity" mentioned above. Professional societies also have great opportunities to address dual-use issues through their annual conferences or training programs; for example, the Wellcome Trust in the UK organized the Summer School on Neuroscience, Ethics and Society 2005 and the European Neuroscience and Society Network organized training programs between 2007 and 2012, including the "Interdisciplinary NeuroSchool" in 2008. Then, the growing number of neuroethics center could facilitate discussion within the community and engage with the security as well as the neuroscience community.

How Can It Be Internationally Promoted?

It has been reported that the implementation of such resources at different academic institutions has been taking place internationally although the number is limited (National Research Council 2010). For the next step, what is really needed to mitigate against a slow advance in the spread and uptake of international biosecurity education is the development of an international network to share emerging best practice.[1] Currently, international efforts to advance education on dual-use issues have only been developed by individual academic institutions, i.e. efforts to date have been mainly "bottom-up." In order to further advance the agenda, "top-down" efforts to provide a structural change in the education culture of the life sciences are also needed. For this purpose, the high-level engagement of governments with coordinated policy decisions is essential.

The Seventh Review Conference of the *BTWC* in December 2011 demonstrated a possible strategic approach.[2] This became a major opportunity for the international community to advance the topic of education as a high priority. Firstly, efforts need to concentrate on the facilitation of the implementation of biosecurity education at different academic institutions. Secondly, efforts are needed to encourage the formalization of further international frameworks to discuss and promote the enhancement of dual-use and biosecurity awareness among scientists. Rappert (2010) has suggested that such formalization may include:

- The establishment of international state coordinators and/or regional coordinators
- The organization of continued international workshops to share best practice
- Provisions of bilateral and multilateral assistance
- Greater incorporation of civil society organizations into the BTWC
- Yearly reporting of activities

By engaging with biosecurity education, neuroscientists can play a significant part in the formation of the next generation of scientists in whose hands the future well-being of society may well be assured.

In result of the Seventh Review Conference, the *BTWC* has set out annual meeting process including Standing Agenda Item – annual agenda item – towards the Eighth Review Conference in 2016. Professional societies, journal editors, and universities in neuroethics field, which were identified in the previous section, should further collaborate with their own national focal points of the BTWC in order to help facilitate possible policy decisions at the international level. In other words there are many opportunities for neuroscientists and neuroethicists to play international roles to promote awareness raising.

[1]Landau Network Centro Volta. http://www.centrovolta.it/landau/2009/12/10/PromotingSustainableEducationAndAwarenessRaisingOnBiosecurityAndDualUse.aspx. Accessed 9 Nov 2012.

[2]Since 1980, Review Conferences have taken place once every 5 years and the next Review Conference will take place in December 2011.

Conclusions and Future Directions

This chapter identified the pervasive lack of awareness and the lack of education of dual-use topics for neuroscientists. The chapter comparatively analyzed how such deficiency could be distinct from the current state of biosecurity education in the broader context of the life sciences. The analysis provided a nearly identical picture in life science fields. However, there has been a series of achievements to develop a comprehensive approach to promote education for life scientists. The term comprehensiveness here characterizes the following points. Efforts are needed to develop a diverse range of education material, training program, and networking. It also requires the engagement of all stakeholders, including science communities, funding agencies, university accreditation bodies, journal editors, export control branches, public health sectors, and security communities. Moreover, actions are necessary to be taken at all levels to change the culture of responsible conduct from the individual level to the international policy world combining bottom-up and top-down efforts. Such comprehensiveness is instrumental to help mitigate the current lack of awareness of dual-use issues in neuroscience community. Indeed, the chapter indicated that there is certain possibility to achieve this endeavor.

Cross-References

▶ Biosecurity as a Normative Challenge
▶ International Legal Restraints on Chemical and Biological Weapons
▶ Neuroethics of Warfare
▶ Weaponization of Neuroscience

References

Atlas, R. M., & Dando, M. R. (2006). The dual-use dilemma for the life sciences: Perspectives, conundrums, and global solutions. *Biosecurity and Bioterrorism, 4*(3), 276–286.

Dando, M. R. (2008). Developing educational modules for life scientists accelerating the process through an open source initiative. Presented to the *IWG–LNCV biological workshop and round table on fostering the biosecurity norm*: *An educational module for life sciences students*, 27 Oct at the Municipality of Como, Italy.

BTWC. (2008). *Report of the meeting of states parties*, 12 December, BWC/MSP/2008/5. Geneva: United Nations.

Dando, M. R., & Rappert, B. (2005). Codes of conduct for the life sciences: Some insights from UK academia. *Bradford briefing papers*, 16 May 2005. http://www.brad.ac.uk/acad/sbtwc/briefing/BP_16_2ndseries.pdf. Accessed 9 Nov 2012.

Feaks, D., Rappert, B., & McLeish, C. (2007). Introduction: A web of prevention? In B. Rappert & C. McLeish (Eds.), *A web of prevention: Biological weapons, life science and the governance of research* (pp. 1–13). London: Earthscan.

Freedman, D. (2010) Israle. In B. Rappert (Ed.), *Education and ethics in the life sciences*: *Strengthening the prohibition of biological weapons* (pp. 75–91). Canberra: Australian National University E Press. http://epress.anu.edu.au/education_ethics.html. Accessed 9 Nov 2012.

Fukushi, T. (2012). Neuroethics and biosecurity in Japan, presented at *an experts meeting on neuroethics at the University of Bradford*, 21 July. Bradford, UK.

Gorski, A., & Spier, R. E. (Eds.)(2010). Special issue section: The advancement of science and the dilemma of dual-use. *Science and Engineering Ethics*, *16*(1), 1–219.

Horgan, J. (2005). The forgotten era of brain chips. *Scientific American, 293*(4), 66–73.

Inter-Academy Panel. (2005). *IAP statement on biosecurity*. http://www.interacademies.net/File.aspx?id=5401. Accessed 9 Nov 2012.

Japan in consultation with JACKSNNZ. (2008). *Oversight*, *education*, *awareness raising*, *and codes of conduct for preventing the misuse of bio-science and biotechnology*, BWC/MSP/2008/MX/WP.21, 14 Aug. http://daccess-ods.un.org/TMP/5950537.32395172.html. Accessed 9 Nov 2012.

Kuhlau, F., Eriksson, S. F., Evers, K., & Hoglund, A. T. (2008). Taking due care: Moral obligations in dual use research. *Bioethics, 22*(9), 477–487.

Lombera, S., Fine, A., Grunau, R. E., & Illes, J. (2010). Ethics in neuroscience graduate training programs: Views and models from Canada. *Mind, Brain, and Education, 4*(1), 20–27.

Mancini, G., & Revill, J. (2008). *Fostering the biosecurity norm*: *Biosecurity education for the next generation of life scientists*. Bradford: University of Bradford. http://www.brad.ac.uk/bioethics/media/SSIS/Bioethics/docs/European_Case_study.pdf. Accessed 9 Nov 2012.

Miller, S., & Selgelid, M. J. (2007). Ethical and philosophical consideration of the dual-use dilemma in the biological sciences. *Science and Engineering Ethics, 13*(4), 523–580.

Millet, P. (2010). The biological weapons convention: Securing biology in the twenty-first century. *Journal of Conflict and Security Law, 15*(1), 25–43.

Minehata, M. (2010). An investigation of biosecurity education for life scientists in the Asia Pacific region. *Research monograph for the Wellcome Trust project on building a sustainable capacity in dual-use bioethics*. Exeter and Bradford: University of Exeter and University of Bradford. http://www.brad.ac.uk/bioethics/media/ssis/bioethics/docs/Asia-Pacific-Biosec-Investigation.pdf. Accessed 9 Nov 2012.

Minehata, M., & Shinomiya, N. (2010). Japan: Obstacles, lesson and future. In B. Rappert (Ed.), *Education and ethics in the life sciences*: *Strengthening the prohibition of biological weapons* (pp. 93–114). Canberra: Australian National University E Press. http://epress.anu.edu.au/education_ethics.html Accessed 9 Nov 2012.

Moreno, J. D. (2012). *Mind wars: Brain science and the military in the twenty-first century* (2nd ed.). New York: Bellevue Literary Press.

National Research Council. (2006). *Globalization*, *biosecurity*, *and the future of the life sciences*. Washington, DC: National Academies. http://www.nap.edu/catalog.php?record_id=11567. Accessed 9 Nov 2012.

National Research Council. (2009). *A survey of attitudes and actions on dual-use research in the life sciences*: *A collaborative effort of the National Research Council and the American Association for the Advancement of Science*. Washington, DC: National Academies. http://www.nap.edu/catalog.php?record_id=12460. Accessed 9 Nov 2012.

National Research Council. (2010). *Challenges and opportunities for education about dual use issues in the life sciences*. Washington, DC: National Academies. http://www.nap.edu/catalog.php?record_id=12958.

Nature. (2003). Statement on the consideration of biodefence and biosecurity. *Nature, 421*(6925), 771.

Netherlands. (2008). *Development of a code of conduct on biosecurity*, BWC/MSP/2008/MX/WP.8, July 30. http://daccess-ods.un.org/TMP/7816224.69425201.html. Accessed 9 Nov 2012.

NSABB. (2008). Strategic Plan for Outreach and Education on Dual Use Research Issues. *Report of the NSABB*, Washington, DC: NSABB. http://oba.od.nih.gov/biosecurity/news_events_oba.html#NSABB. Accessed 9 Nov 2012.

OECD. (2007). *Best practice guidelines on biosecurity for BRCS*. Paris: OECD.

Rappert, B. (2010). An action plan for education: Possibilities and plans. Presented at the *ESRC-JSPS collaborative seminar*: *Dual-use education for life scientists*: *Mapping the current global landscape and developments*. 15–16 July, UK: University of Bradford. http://www.brad.ac.uk/acad/sbtwc/dube/resource/ESRC_seminar_web/ppt/Rappert_ActionPLan.pdf. Accessed 9 Nov 2012.

Rappert, B., Chevrier, M. I., Dando, M. R. (2006). In-depth implementation of the BTWC: Education and outreach. *Bradford review conference papers*, 18. http://www.brad.ac.uk/acad/sbtwc/briefing/RCP_18.pdf. Accessed 9 Nov 2012.

Revill, J., Mancini, G., Minehata, M., Shinomiya, N. (2009). Biosecurity education: Surveys from Europe and Japan. *Background paper for the international workshop on promoting education on dual-use issues in the life sciences*. 16–18 Nov, Warsaw, Poland: Polish Academy of Sciences.

Sahakian, B. J., & Morein-Zamir, S. (2009). Neuroscientists need neuroethics training. *Science, 325*, 147.

Schliermeier, Q. (2002). Hostage deaths put gas weapons in spotlight. *Nature, 420*, 7.

Selgelid, M. J. (2007). A tale of two studies: Ethics, bioterrorism, and the censorship of science. *The Hastings Centre Report, 37*(3), 35–43.

Smith, G., Davison, N., Koppelman, B. (2010). The role of scientists in assessing the risks of dual-use research in the life sciences. In J. L. Finney & I. Slaus (Eds.) *Assessing the threat of weapons of destruction*: *The role of independent scientists* (pp. 137–140), (NATO Science for Peace and Security Series E: Human and Societal Dynamics – Vol. 61), Amsterdam: IOP Press.

Snyder, P. J. (2009). Delgado's brave bulls: The marketing of a seductive idea and a lesson for contemporary biomedical research. In P. J. Snyder (Ed.), *Science and the media: Delgado's brave bulls and the ethics of scientific disclosure* (pp. 25–40). London: Elsevier.

Society for Neuroscience. (2012), *Guidelines for authors of scientific communications*, http://www.sfn.org/index.aspx?pagename=responsibleConduct_authorsOfResearchManuscripts. Accessed 13 Nov 2012.

Somerville, M. A., & Atlas, R. M. (2005). Ethics: A weapon to counter bioterrorism. *Science, 307*(5717), 1881–1882.

Streatfeild, D. (2006). *Brainwash: The secret history of mind control*. London: Hodder & Stoughton.

Walther, G. (2012). Ethics in neuroscience curricula: A survey of Australia, Canada, Germany, the UK, and the US. *Neuroethics*. [forthcoming]. http://link.springer.com/article/10.1007/s12152-012-9168-2. Accessed 13 Nov 2012.

World Health Organization. (2007). *Scientific working group on life science research and global health security*: *Report of the first meeting*, 16–18 Oct 2006, Geneva: WHO. http://www.who.int/entity/csr/resources/publications/deliberate/WHO_CDS_EPR_2007_4n.pdf. Accessed 9 Nov 2012.

Neuroscience Advances and Future Warfare

114

Malcolm Dando

Contents

Abstract

This paper begins by recalling that advances in neuroscience were used for hostile purposes, for example, in the development of lethal nerve gasses, in the last century, and it is argued that in the kinds of asymmetric warfare likely to characterize coming decades, such advances could again be utilized to develop novel weapons. The paper then suggests that the idea that the problem is that bioterrorists will immediately be able to design and use advanced biological and chemical weapons is misguided and that the real question is how the wholesale militarization of the life sciences can be prevented. It is in that context that the paper examines the dangers of misuse that could arise from some current developments in neuroscience. It is argued, for example, that benignly intended civil work on transcranial magnetic stimulation (TMS) and brain-computer interfaces (BCIs) has to be understood in the context of modern military interests in data collection and analysis from drones and the probable development of

M. Dando
Division of Peace Studies, School of Social and International Studies, University of Bradford, Bradford, UK
e-mail: mrdando@bradford.ac.uk; m.r.dando@bradford.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_139,

autonomously acting systems. The difficulties that such novel weapon-related developments will cause for our present understanding of morality and international law are reviewed, and finally, it is suggested that neuroscientists trying to adjust their concepts of responsible conduct in these circumstances will need the help of neuroethicists.

Introduction

In 2006 the *Journal of the History of the Neurosciences* carried a long paper (Schmaltz 2006) titled "Neurosciences and Research on Chemical Weapons of Mass Destruction in Nazi Germany." The paper gives a detailed account, based on the most recently available documents, of how work on potential pesticides led to the discovery of tabun, sarin, and eventually Soman – the first of the lethal nerve gasses – in the context of the force-on-force mechanized warfare of that time.

What is striking about this account, as is clear from the title, is the way in which basic neuroscience research was "enmeshed" in applied offensive and defensive chemical warfare research. The primary effects of the nerve agents, for example, "were originally identified at a branch laboratory of the Military Medical Academy, by Gremels, the Director of the Pharmacological Institute of the University of Marburgh," and "The methods and models applied at the KWI [Kaiser Wilhelm Institute] for Medical Research, originally developed for research on vitamins and enzymes, were an important condition for the1944 discovery of the nerve agent Soman" which was at that time the most potent inhibitor of cholinesterase and thus the most deadly nerve agent.

Such historical research should alert us to the need to be careful that the very rapid advances in neuroscience being made in our time are not similarly used for warfare and other hostile purposes. In that regard, we have to acknowledge that in the new kinds of asymmetric warfare that will characterize coming decades (Development, Concepts and Doctrine Centre 2010):

> The CBRN [Chemical, Biological, Radiological, Nuclear] threat from state and non-state actors is likely to increase, facilitated by lowering of some entry barriers, dual-purpose industrial facilities and the proliferation of technical knowledge and expertise...

Moreover, this forecast continues: "Terrorist attacks using chemical, biological and radiological weapons are likely, as are mass-casualty attacks using novel methods." It is worth noting that in this estimate the word "likely" is stated to indicate a 60–90 % probability.

We have, of course, become familiar, since the turn of the century, with the fact that the genomics revolution and neuroimaging technologies will transform our understanding of the function of neurotransmitters and neuroreceptors in critical brain circuits (Andreasen 2001) and that this knowledge might be misused. What is clear from more recent accounts of the possible advances in neuroscience is the importance of information technology more generally in understanding and

manipulating brain functions (Horstman 2010), and it cannot be ignored that information technology is becoming more and more important in the operations of military forces in many countries.

The aim of this paper is not to attempt a survey of all the kinds of neuroscience that could perhaps be misused for hostile purposes such as the epigenetic influences of the social environment (Champagne and Curley 2011), but rather, by discussing some more obvious examples, to illustrate why considerations of the darker side of possible advances and applications are important for neuroscientists and, by implication, neuroethicists, to consider. It should also be noted that the concentration here is on uses in acts of warfare rather than in other possible misuses such as terror, interrogation, or torture.

Framing the Question

In the report of the Commission on the Prevention of WMD Proliferation and Terrorism in 2008, US Senator Bob Graham and his colleagues concluded that (Commission on Prevention of WMD Proliferation and Terrorism 2008):

> . . .it is more likely than not that a weapon of mass destruction will be used in a terrorist attack somewhere in the world by the end of 2013.

Immediately following that statement, the Commission added:

> The Commission further believes that terrorists are more likely to be able to obtain and use a biological weapon than a nuclear weapon. . .

In short, a biological weapon would be used, as a weapon of mass destruction by terrorists by the end of 2013.

That has always seemed a wildly exaggerated threat assessment to me. It has been clear for at least 50 years that biological weapons could be used to cause mass casualties (Wheelis et al. 2006), but weaponization of biological agents is clearly difficult unless a State decides to embark on an offensive biological weapons program. Indeed, the 2001 anthrax letter attacks in the USA used material developed in the 25-year-long US offensive program, and that is probably the only such attack that has taken place since the Second World War. The idea that terrorists would be easily able to utilize advanced techniques of modern biology to produce and use new types of dangerous biological weapons without access to the resources of a State seems even more farfetched as the tacit knowledge required would not likely be available to nonspecialists just reading the scientific literature (Vogel 2008). These facts strongly suggest that there is something amiss with the overconcentration on the threat of bioterrorism that we have witnessed in the last 10–15 years.

A further worry arises when the overconcentration on bioterrorism is linked uncritically to the threat of modern biological research being of "dual-use concern" and therefore needing new forms of governmental oversight

(Gottron and Shea 2012). If the discussion in 2005 over the publication of the sequencing and synthesis of the genome of Spanish influenza did not suggest that in a vanishing small number of instances would consensus be reached that the risks would outweigh the benefits of publication, surely the recent debate on the creation of mammalian-transmissible H5N1 influenza should have given pause for thought (Novossiolova et al. 2012). Do we really need to have a vast oversight system when even the prevention of the publication of a route to a "perfect bioweapon" is not thought necessary?

One possible response to this line of argument is to reject the whole idea of being concerned about the hostile misuse of the modern life sciences. As one commentary on the H5N1 debate suggested (Trevan 2012):

> The best strategy to stop biological attacks is to make biological weapons unattractive by making preparations and response so effective that the consequences are no worse than a train wreck. Increases understanding of transmissibility and pathogenicity will enable countries to identify threats earlier, develop better vaccines, produce them more quickly and develop broad-spectrum defences to diseases. . .

That again seems an unrealistic approach to me, in part, because I find it difficult to imagine States devoting sufficient resources to achieving such ends.

More fundamentally, however, I think it is a wildly optimistic forecast of what is likely to happen if a State-level offensive biological arms race is allowed to develop in coming decades on the back of the revolution in modern biology and its increasing industrial application. A much better way to frame our question and to answer it properly is not in terms of potential bioterrorism and response, but to see that as a small part of a much larger future danger. That danger was set out very clearly by Professor Mathew Meselson of Harvard University in 2000 (Meselson 2000):

> . . .During the century ahead, as our ability to modify fundamental life processes continues its rapid advance, we will be able not only to devise additional ways to destroy life but will also become able to manipulate it - including the processes of cognition, development, reproduction, and inheritance. . .

Meselson then pointed out the consequences if human society allowed this to happen:

> . . .A world in which these capabilities are widely employed for hostile purposes would be a world in which the very nature of conflict had radically changed. Therein could lie unprecedented opportunities for violence, coercion, repression, or subjugation. . .

Moreover, these capabilities would be widespread:

> . . .Unlike the technologies of conventional or even nuclear weapons, biotechnology has the potential to place mass destructive capabilities in a multitude of hands and, in coming decades, to reach deeply into what we are and how we regard ourselves. It should be evident that any intensive exploitation of biotechnology for hostile purposes could take humanity down a particularly undesirable path.

Therefore, while I believe that bioterrorism is given too much attention at the present time, I would certainly not discount it in a future in which the use of

modern biology for hostile purposes has become a commonplace among States. A major danger at the present time is precisely in regard to the activities of some States which, as discussed in the Royal Society Brain Waves Module 3 report on *Neuroscience, conflict and security* (Royal Society 2012), appear to be endangering the prohibition regime because of their interest in new forms of incapacitating chemical agents derived from ongoing advances in neuroscience.

An idea of what might well happen in a prolonged arms race in biological weapons was set out by three US analysts in 2003 (Petro et al. 2003). They suggested that there would likely be three overlapping phase. The first would involve the use of traditional agents such as anthrax, but as there are only a limited number of such agents, the defense would eventually be able to catch up. Then the offense would move to modify such agents by the use of genetic engineering, but, again, as there are only a limited number of such modifications possible, in theory the defense could again catch up. However, as our understanding of living systems continues to improve, there would be many more targets that could be attacked in an increasing number of diverse ways through the use of new advanced biological agents. In such a situation, I find it difficult to see anything other than a prolonged period of offensive superiority even if huge resources were to be thrown into biodefense.

Of course, in such an environment, it would be difficult to ensure that others understood that biodefense was not, in fact, part of an offensive program. Additionally, efforts to "enhance" military forces, as has been envisioned in some recent literature, might, again, be perceived as degradation of another State's "normal" forces. Moreover, benignly intended uses of modern technology within the military could have perverse effects by what has been termed "reverse dual use" within civil society (Marchant and Gulley 2010). Indeed it is necessary to emphasize that the problem is the potential for multiple means of misuse over coming decades not of terrorists taking up advanced biotechnology to create weapons of mass destruction now. I will not pursue these complexities further here except to say that we have to be aware that in such muddy waters, there can be subtle marketing of new weapons systems – such as the misappropriation of humanitarian arguments for new "nonlethal" weapons (Tafolla et al. 2012) in the same way as new chemical weapons have been justified in the past. This is particularly important because at the moment neuroscientists, like most practicing life scientists, have very little awareness of the potential for their benignly intended work to be misused for hostile purposes (Australia et al. 2011).

Novel Neurotechnologies

Recently some neuroethicists have begun to analyze the problem of dual use. In the UK, for example, the Nuffield Council on Bioethics has decided to investigate three novel brain technology areas: brain-computer interfaces, neurostimulation, and neural stem cell therapy. In its *Consultation Paper* of March 2012 (Nuffield Council

on Bioethics 2012), the council clearly notes that there are possible concerns in regard to the military/hostile uses of brain-computer interface technologies (BCIs) and of transcranial magnetic stimulation (TMS).

TMS

The potential misuse of TMS has been examined recently in a collection of essays on *Innovation, Dual Use and Security: Managing Risks of Emerging Biological and Chemical Technologies* (Tucker 2012). The authors of the essays were asked to examine the present state and future trajectories of a series of different technologies, then to assess the risks that the technologies might pose and finally the governance strategies that might be applied to minimize the risks. In his introduction the editor made it obvious that the frame of analysis was dual use and bioterrorism with subsections on defining dual use, the dual-use landscape (past and present), the rise of synthetic genomics, technical hurdles to bioterrorism, and so on.

The chapter on TMS was written by Jonathan Moreno and shows that while the mechanism of action is not understood (Moreno 2012a):

> . . .Over the past twenty years, repetitive TMS has evolved from an experimental diagnostic technique into a mature technology that has been approved for treatment of major depressive disorder and has other promising clinical applications.

However, the production of the equipment and diffusion of the technology remains relatively limited:

> . . .The number of vendors of TMS equipment today is limited, although interest in the therapeutic use of the technology is growing amongst advanced industrial countries.... Once the safety and efficacy of TMS have been demonstrated, it is likely to spread to additional countries.

So should there be potential for misuse of TMS, it is not yet impossible to put in place sensible control mechanisms such as those being developed in regard to aspects of synthetic biology (Garfinkel et al. 2007).

TMS is not in the category of a technology that could be misused to create a weapon of mass destruction (WMD), but the author of the essay, who has given considerable thought to the misuse of neurotechnologies (Moreno 2012b), lists a series of possible concerns. In his view:

> Repetitive TMS is a clear case of a dual-use technology.... Given the limited expertise needed to employ the technique. . . states or terrorist organisations could potentially misuse it for harmful purposes....

He lists, for example, “erasing” the memory of someone to make him unable to disclose information under interrogation or to “enhance” the ability of a terrorist to carry out an attack.

Moreno also cites the interest of the US military in TMS and its possible deployment in the field within one or two decades to enhance “top-down visuospatial attention” in combination with fMRI and this perhaps leading to an

enhancement arms race in which soldiers have to accept increasing amounts of cognitive modification. The reason for the military interest in technologies like TMS was set out in an article by a team with a member of the US Air Force Research Laboratory as the lead author (McKinley et al. 2012):

> . . .These techniques are perhaps best suited for career fields where certain cognitive skills such as vigilance and threat detection are essential in preserving human life. Because such jobs are plentiful in the military, it is no surprise that the US Air Force has recently begun investing in noninvasive brain stimulation for its efficacy in benefiting human cognitive performance....

Perhaps of more concern, Moreno cites recent work (Young et al. 2010) which suggests that TMS might be misapplied to manipulate moral judgements or beliefs about right and wrong, it seems, by elevating consideration of the outcome of another's actions above their intent, through disruption of the right temporoparietal junction.

What is of particular interest is the discussion of the options for governance. It should be noted here that the author of a study of the promises and perils of noninvasive brain stimulation (Heinrichs 2012) to my mind appears to totally misunderstand the problem of dual use when he concludes that "most issues in medical and research ethics such as . . .potential for misuse can be handled with manageable effort" as that is precisely what cannot be achieved in regard to dual use and why there has been so much concern. The problem is that benignly intended work, carried out quite properly according to accepted required standards, might later be misused by others. In his chapter Moreno suggests two different options. First, since there are still a limited number of manufacturers, it should be possible to track international sales (even if it is difficult to be sure of how the equipment is used after the sale). Second soft-law and informal approaches such as awareness-raising of possible misuse among users and the development of education and codes of conduct might be developed. The final chapter of the book (Bansak and Tucker 2012) sets out the assessment of the risks of misuse and of the governance possibilities for each technology in tabular form. Then in a final table, the risk of misuse of each technology is compared with its potential governance. It is worth quoting the conclusion in full in regard to TMS:

> The technology in the lower right-hand corner of the matrix, transcranial magnetic stimulation (TMS), poses a low risk of misuse because it can be applied to only one individual at a time rather than to a large group or population. For this reason, the potential misuse of TMS to support coercive interrogation or to erase memories is more of a human rights concern than a potential threat to national security. . . .

And the text continues to assert that:

> . . . At the same time, TMS has a high governability score because its clinical use is heavily regulated on safety and ethical grounds. That TMS consists primarily of hardware also makes it more susceptible to governance. Because the risk of misuse of TMS is low, the technology does not warrant the development of governance strategies.

Dealing first with the question of the numbers of people who might be affected, the article quoted previously (with the member of the US Air Force Research

Laboratory as lead author) makes clear the current demands for data analysis in Iraq and Afghanistan (McKinley et al. 2012):

> . . .Each combat air patrol (CAP) currently consists of 4RPA [Remotely Piloted Aircraft], 43 personnel for mission control, 59 for launch and recovery and 66 for processing, exploitation and dissemination. Not surprisingly, the strain on manpower is one of the major challenges facing the Air Force in coming years....

So to accept the argument for ignoring governance of TMS, we have first to set aside the prospect of thousands of soldiers (including conscripts) spending their working days subject to a technology of unknown long-term health consequences. The main point, however, is that on this reading the government should not spend any effort – not even awareness-raising and education on dual use and biosecurity – in regard to TMS. The dual use/bioterrorism puts the focus elsewhere.

On the other hand, taking the view that bioterrorism is not going to be a major problem for some time and that the real problem we face is preventing the wholesale militarization of the life sciences suggests a quite different answer. Particularly, given the difficulty of controlling technologies once they have been incorporated into military equipment and operations, should we not, instead of ignoring TMS, take the opportunity provided by the possibility of its governance to go all out to get it under control now? At least attempting to reach that objective would provide the neuroscience, neuroethics, and policy-making communities experience and models that could help in the future as the revolution in the life sciences and more difficult problems related to militarization and potential hostile misuse develop.

BCIs

As the Nuffield Council's consultation paper explained, in a BCI (Nuffield Council on Bioethics 2012):

> Patterns of electrical activity in the brain that accompany an individual's thoughts and intentions give rise to characteristic brain signals which can be detected.... These electrical signals can be analysed and converted, with the help of a computer, into various commands and appropriate actions....

The paper goes on to note a number of BCI applications that are known to be of interest to the military such as augmentation of muscle strength for the individual soldier and to enable remote control of vehicles. What is of particular interest here is the view expressed that (Nuffield Council on Bioethics 2012):

> . . .Finally, 'telepresence' is where a soldier, whilst physically present elsewhere, has the ability to sense and interact in a removed and real world location, such as with a demolition robot or unmanned vehicle (drone) through a BCI connection.

So the council has decided to examine two technologies that are of increasing importance to the military: TMS, as we have seen, is a possible means of

enhancing data analysis from remote sources, and BCIs could be important in the future as part of a control system for the operation of drones. The association of these two noninvasive neurotechnologies with advances in information technology (IT) is crucial to grasp because developments in IT are critical for military forces at the present time as they seek to gain technological superiority in this field. Thus it can reasonably be argued that the council is carrying out a study of one of the current cutting edges in the militarization of neuroscience.

One recent US analysis considers the impact of IT on military superiority in the future. IT is viewed in this analysis as an exponentially growing technology as it is one in which "performance per dollar multiplies over time." Such exponential growth in military-related technologies can produce opportunities (Kopp 2012):

> . . .to devise entirely new solutions to longstanding, or entirely new problems. . .[these] can frequently produce highly disruptive effect, as the new solution will often exploit systemic weaknesses in the opponent's capabilities that cannot be easily overcome by established means.

Given that IT is available commercially around the world, the paper suggests that all players are on a level playing field, and in such circumstances "the player who can best exploit talent to an advantage - all else being equal - will inevitably win."

It is in that general context and the increasing use of remotely piloted aircraft, drones or unmanned aircraft systems (UASs) that the potential use of TMS and BCIs in such equipment has to be considered. This is not a theoretical problem, as concerns about the use of drones for assassination purposes among international lawyers amply illustrate (Bowcott 2012).

Any doubt as to the potential involvement of BCIs in the long-distance control of UASs is easily dispelled by a minimal review of the literature that goes beyond simple extension of human capabilities like robotic arms (Brunner et al. 2011). For example, a paper published in 2011, reported utilization of a noninvasive BCI system employing EEG recordings of sensorimotor rhythms (SMRs) to demonstrate human capabilities to exert three-dimensional control of a virtual helicopter. The paper suggested that (Doud et al. 2011):

> . . .three-dimensional movement of a virtual helicopter. . .is fast, accurate, and continuous. In this system, the virtual helicopter's forward-backward translation and elevation controls were activated through the modulation of the sensorimotor rhythms that were converted to forces applied to the virtual helicopter at every simulation time step, and the helicopter's angle of left or right rotation was linearly mapped, with high resolution, from sensorimotor rhythms associated with other motor imagination. . . .

It is fanciful to imagine that such work is not of interest to the military even if there is no indication of military funding for this particular publication.

The UK Ministry of Defence (MOD) has certainly taken the moral, legal, and ethical issues involved in the operation of UASs seriously, devoting a separate chapter to discuss them in its Joint Doctrine Note 2/11 on *The UK Approach to*

Unmanned Aircraft Systems (Ministry of Defence 2011). In this publication careful differentiation is made between an unmanned aircraft and an unmanned aircraft system. The latter is defined as:

> . . .a system, whose components include the unmanned aircraft and all equipment, network and personnel necessary to control the unmanned aircraft.

All operations by UASs are governed by the Law of Armed Conflict (LOAC):

> . . .Firstly, weapons law guides whether a weapon and its generic uses are lawful; secondly, targeting law determines whether the use of a particular weapon is lawful on a specific mission or in specific circumstances....

This raises questions such as whether a remote pilot has adequate knowledge to make reasoned judgements about proportionality and discrimination. The law of armed conflict (International Humanitarian Law, IHL) is based on the concept of legal responsibility. As the MOD paper notes, "[L]egal responsibility for any military activity remains with the last person to issue the command authorising a specific activity."

Also of interest here is also the legal status of the remote operator. In a section on "The Remote Warrior," the paper asks a series of difficult questions:

> . . .With kinetic operations being controlled from several thousand miles away, such as those in Afghanistan currently being conducted from the continental US, LOAC issues are further complicated. The concept of *fighting from barracks* as it has been termed raises a number of interesting areas for debate....

These questions, as the MOD paper sets them out, include:

> . . .Is the *Reaper* [UAS] operator walking the streets of his home town after a shift a legitimate target as a combatant? Would an attack by a Taliban sympathiser be an act of war under international law. . .? Does a person who has the right to kill as a combatant while in a control cabin cease to be a combatant that evening on his way home?....

Going beyond the legal issues, the paper raises a number of what it sees as moral and ethical issues such as whether the capability to wage war from a distance with no risk to our soldiers' lives risks unnecessary escalation of conflicts and brings the legitimacy of our actions into question.

Autonomous Weapons

What is most striking about the MOD paper, however, is the constant concern about the likely coming of autonomous systems and the implications, legal and ethical, of that future. For example, the paper notes the increasing complexity and pace of modern warfare:

> . . .There is also an increasing body of discussion that suggests that the increasing speed, confusion and information overload of modern war may make human response inadequate and that the environment will be '*too complex for a human to direct*' and this has already been exemplified by the adoption. . .of autonomous weapon systems. . . .

Such considerations raise more difficult questions about responsibility:

> . . . Is a programmer guilty of a war crime if a system error leads to an illegal act? Where is the intent required for an accident to become a crime?

And the paper considers that there is an urgent need for policy development:

> . . . The pace of technological development is accelerating and the UK must establish quickly a clear policy on what will constitute acceptable machine behaviour in the future; there is already a significant body of scientific opinion that believes in banning autonomous weapons outright, countered by an acceptance in other areas that autonomy is inevitable....

The net result of the analysis in this chapter of the MOD paper is bleak:

> . . .There is a danger that time is running out - is debate and development of policy even still possible, or is the technological genie already out of the bottle, embarking us all on an incremental and involuntary journey towards a *Terminator*-like reality?

My own answer to that question is probably yes, but not necessarily, as the implementation of a new technology is not preordained but contingent on social processes (Price 1997) and there are certainly efforts under way, for example, by the International Committee of the Red Cross to avoid such an outcome. The ICRC is certainly clear about the importance of dealing with the question of autonomous weapons (ICRC 2011):

> An autonomous weapon system is one that can learn or adapt its functioning in response to changing circumstances in the environment in which it is deployed. A truly autonomous system would have artificial intelligence that would have to be capable of implementing IHL. Such systems have not yet been weaponized although there is considerable interest within expert literature and considerable funding of relevant research. The deployment of such systems would reflect a paradigm shift and a major qualitative change in the conduct of hostilities. It also raises a range of fundamental legal, ethical and societal issues which need to be considered *before* such systems are developed or deployed.

The difficulties of dealing with autonomous weapons, however, were recently exemplified by the failure even to get these weapons into the draft of the Arms Trade Treaty (Bolton 2012).

Moral Machines?

Although the work of neuroscientists could be seen as helping to take us down the road to autonomous weapons, it could be argued that once such systems are in place, there is little point in asking more questions of neuroscientists. However, that is to ignore the fact that their technologies are also part of the investigatory apparatus of neuroscience that is being used to research moral decision-making. Moreover, neuroscience may also be involved in further advances in which new information technologies materials and methods are derived from our growing understanding of the operations of the nervous system (see, e.g., the US DARPA projects on NeuroVision 2 and Systems of Neuromorphic Adaptive Plastic Scalable Electronics – SyNAPSE – as described on the Defence Sciences

Office section of the DARPA website www.darpa.mil). Therefore, it seems to me that neuroscientists will have a continuing responsibility to help guard their work from misuse.

The problem caused by the progress of modern information technology is that in warfare there is a deluge of new data from sources such as drones. TMS may be useful in helping soldiers to handle such data, and BCIs may be useful in reducing the time needed to respond to the data input, but eventually there will be limits to such solutions so that it becomes inevitable on this logic that autonomous weapon systems will have to be developed and deployed. These can be made to deal with larger amounts of data and never get tired, but as the MOD paper points out (Ministry of Defence 2011):

> . . .To a robotic system, a school bus and a tank are the same - merely algorithms in a programme - and the engagement of a target is a singular action; the robot has no sense of ends, ways and means, no need to know **why** it is engaging a target. . . .

For some the solution to that problem is to design moral machines that can apply the Law of Armed Conflict better and more consistently than human beings.

Ronald Arkin, a US engineer who takes his responsibilities for the social impacts of his work very seriously (Bookman 2011), has argued, for example, that (Arkin et al. 2012):

> Weaponised military robots are now a reality. While in general a human remains in the loop for decision making regarding the deployment of lethal force, the trend is clear that targeting and engagement decisions are being moved forward onto these machines as the science of autonomy progresses. . ..

He therefore has put forward:

> . . .the research hypothesis that autonomous systems could ultimately operate more humanely than human warfighters are able to. As part of the research to test this thesis funded by the Army Research Office, an ethical architecture for an autonomous system was developed with the intent of enforcing the principles derived from LOW [Laws of War], thus having the goal of enhancing noncombatant safety and survivability. . ..

Others have argued that ethical judgements are complex and unlikely to be easily programmed into a machine (Prichard 2012) – even if the data required would be available to the machine in the “fog” of war.

Whether moral machines can be built is an empirical issue that will be decided later, but what is not in doubt is that the work of neuroscientists will be utilized in the process of deciding (Arkin et al. 2012):

> . . .researchers should utilize biologically relevant models of ethical behaviour as the basis of research: drawing heavily on ethology, neuroscience, cognitive psychology, and sociology as appropriate as a basis. . ..

Moreover, as progress is made in understanding of how we make moral decisions, it is surely likely that modification of ethical behavior in civil society through means such as TMS will come to be increasingly discussed (Heinzelmann et al. 2012).

Conclusion

People in the military/security field have remarked to me that we civilians have difficulty looking through "the dark side of the telescope." We do not naturally think about people intending to do harm, or as Andy Stirling put it in his essay on the governance of neuroscience for the Royal Society Brain Waves Module 1 (Stirling 2011):

> . . .Although readily foreseeable in the same terms as benign uses, malign applications are typically underestimated in regulatory assessment. . .. Yet easily anticipated effects may be of a magnitude that seriously jeopardises overall benefits. This is exemplified by the paradox that military aims are at the same time so prominent and so under-scrutinised in global research. Roughly one third of worldwide human effort in research and innovation is devoted – directly or indirectly – to refining ways to perpetrate premeditated organised violence. . ..

This section of the essay is subtitled "See no evil."

I therefore took it that my task in this paper was to try to give some idea of what the advances in neuroscience might look like from the biosecurity "dark" side in the hope that this could convey some of the issues that would be considered important to the security community and those, like myself, who hope to develop better means to restrict violence.

In an important multiauthor paper published in *The Columbia Science and Technology Law Review* (Marchant et al. 2012), the rise of autonomous military robotics is outlined and the ethical and policy issues are described. Then currently available governance mechanisms are reviewed, and it is concluded that "contemporary governance architecture regarding the innovation and use of military robots would appear wholly inadequate to the task." Finally, the range of possible future governance mechanisms, from legally binding "hard-law" to "soft-law" approaches, such as codes of conduct, are discussed. The authors do not suggest which approach should be used but call "for a national and international dialogue on appropriate governance of such systems *before* they are deployed." I agree, but note that at present most neuroscientists would not be able to contribute to the dialogue and would probably not even know that it was taking place.

Consequently, I would urge the support of the recommendation of the Royal Society Brain Waves Module 3 on *Neuroscience, conflict and security* in its first recommendation that (Royal Society 2012):

> There needs to be a fresh effort by the appropriate professional bodies to inculcate the awareness of the dual-use challenge (i.e., knowledge and technologies used for beneficial purposes can also be misused for harmful purposes) among neuroscientists at an early stage of their training.

Moreover, the UK has made clear that it supports this approach strongly (United Kingdom 2012):

> . . .The UK shares the view that it is important to look at how the issue of dual-use can be assimilated with broader professional training for scientists in the university curricula in a holistic and sustainable manner both at home and abroad. . ..

As Michael Tennison and Jonathan Moreno concluded in their recent review of the state of the art in the field of neuroscience, ethics, and national security if neuroscientists were aware and educated (Tennison and Moreno 2012):

> . . .Just as many nuclear scientists opposed the development of atomic weapons, contributing to the test ban treaties of the 1960s and the drawdown of armed missiles in the 1980s. . .neuroscientists could consider and promulgate their perspectives on the military implications and ethical issues associated with their work.

As has been cogently argued, it would be best to base such education on the ideas of Responsible Conduct of Research that scientists already have begun to assimilate (Carlson and Frankel 2011) by adding consideration of "external" aspects such as dual use to the "internal" aspects such as research misconduct. Nevertheless, it surely will be necessary to supplement that with at least a basic framework of ethics if scientists are going to contribute effectively to protecting their benignly intended work from hostile misuse.

Cross-References

▶ Biosecurity as a Normative Challenge
▶ Biosecurity Education and Awareness in Neuroscience
▶ Brain Research on Morality and Cognition
▶ Ethics of Brain–Computer Interfaces for Enhancement Purposes
▶ International Legal Restraints on Chemical and Biological Weapons
▶ Neuroethics of Warfare
▶ Weaponization of Neuroscience

References

Andreasen, N. C. (2001). *Brave new brain: Conquering mental illness in the era of the genome*. Oxford: Oxford University Press.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE, 100*(3), 571–589.

Australia, Canada, Japan, New Zealand, Republic of Korea and Switzerland (on behalf of the "JACKSNNZ") and Kenya, Sweden, Ukraine, the United Kingdom of Great Britain and Northern Ireland and the United States of America. (2011). *Revised: Possible approaches to education and awareness-raising among life scientists*. BWC/CONF.VII/WP.20/Rev.1, Geneva: United Nations.

Bansak, K. C., & Tucker, J. B. (2012). Governance of emerging dual-use technologies. In: J. B. Tucker (Ed.), *Tucker* (pp. 305–339). Cambridge, MA: MIT Press

Bolton, M. (2012) A draft treaty? The holes in the draft arms trade treaty. *Global Policy Journal*. http//: www.globalpolicyjournal.com/blog/draft-treaty

Bookman, T. (2011). Governing lethal behavior in robots: T&S interview with Ronald C. Arkin. *IEEE Technology and Society Magazine, 30*(4), 7–11.

Bowcott, O. (2012). US drone attacks threaten 60 years of international law, says UN official. *The Guardian*, p. 26.

Brunner, P., Bianchi, L., Guger, C., Cincotti, F., & Schalk, G. (2011). Current trends in hardware and software for brain-computer interfaces (BCIs). *Journal of Neural Engineering, 8*, 025001.

Carlson, R., & Frankel, M. S. (2011). Reshaping responsible conduct of research education. *Professional Ethics Report, XXIV*(1), 1–3.

Commission on the Prevention of WMD Proliferation and Terrorism. (2008). *World at risk*. New York: Vintage Books/Random House.

Champagne, F. A., & Curley, J. P. (2011). Epigenetic influence of the social environment. Chapter 10. In A. Petronis & J. Mill (Eds.), *Brain, behaviour and epigenetics*. Berlin: Springer.

Development, Concepts and Doctrine Centre. (2010). *Global strategic trends – Out to 2040*. London: Ministry of Defence.

Doud, A. J., Luca, J. P., Pisansky, M. T., & Bin, H. (2011). Continuous three-dimensional control of a virtual helicopter using a motor imagery based brain-computer interface. *PLoS One, 6*(10), e26322.

Garfinkel, M., Endy, D., Epstein, G. L., & Friedman, R. M. (2007). *Synthetic genomics: Options for governance*. Washington, DC: CSIS.

Gottron, F., & Shea, D. A. (2012) *Publishing scientific papers with potential security risks: Issues for Congress*. Congressional Research Service, 7-5700. Washington, DC: US Congress.

Heinrichs, J.-H. (2012). The promise and perils of non-invasive brain stimulation. *International Journal of Law and Psychiatry, 35*, 121–129.

Horstman, J. (2010). *The scientific American brave new brain*. San Francisco: Jossey-Boss.

Heinzelmann, N., Ugazio, G., & Nobler, P. N. (2012). Practical implications of empirically studying moral decision-making. *Frontiers in Neuroscience, 6*(Article 94), 1–14.

ICRC. (2011). International Humanitarian Law and the challenges of contemporary armed conflicts. 31IC/11/5.1.2, 31st International Conference of the Red Cross and Red Crescent, Geneva.

Kopp, C. (2012). Technological strategy in the age of exponential growth. *JFQ, 66*(3), 42–47.

Marchant, G. E., et al. (2012). International governance of autonomous military robots. *The Columbia Science and Technology Law Review, XII*, 272–315.

Marchant, G., & Gulley, L. (2010). National security neuroscience and the reverse dual-use dilemma. *AJOB Neuroscience, 1*(2), 20–22.

McKinley, R. A., Bridges, N., Walters, C. M., & Nelson, J. (2012). Modulating the brain at work using noninvasive transcranial stimulation. *NeuroImage, 59*, 129–137.

Meselson, M. (2000). Averting the hostile exploitation of biotechnology. *The Chemical and Biological Weapons Conventions Bulletin, 48*, 16–19.

Ministry of Defence. (2011). *The UK Approach to Unmanned Aircraft Systems*. Joint Doctrine Note 2/11. London: Ministry of Defence.

Moreno, J. D. (2012a). Transcranial magnetic stimulation, In: J. B. Tucker (Ed.), *Tucker* (pp. 223–222). Cambridge, MA: MIT Press

Moreno, J. D. (2012b). *Mind wars: Brain science and the military in the 21st century* (2nd ed.). New York: Bellevue Library Press.

Novossiolova, T., Minehata, M., & Dando, M. R. (2012). The creation of a contagious H5N1 Influenza virus: Implications for the education of life scientists. *Journal of Terrorism Research*, *3*(1), Special Issue – Assessing the Emergency Response to Terrorism: Novossiolova. http://www.ojs.st-andrews.ac.uk/index.php/jtr/article/view/417

Nuffield Council on Bioethics. (2012). *Consultation: Novel neurotechnologies: intervening in the brain*. London: Nuffield Council on Bioethics.

Petro, J. B., Plasse, T. R., & McNulty, J. A. (2003). Biotechnology: Impact on biological warfare and biodefense. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science, 1*(3), 161–169.

Price, R. M. (1997). *The chemical weapons taboo*. Ithica: Cornell University Press.

Prichard, M. S. (2012). Moral machines? *Science and Engineering Ethics, 18*, 411–417.

Royal Society. (2012). *Neuroscience, conflict and security. Brain waves module 3*. London: Royal Society.

Schmaltz, F. (2006). Neurosciences and research on chemical weapons of mass destruction in Nazi Germany. *Journal of the History of the Neurosciences, 15*, 186–209.

Stirling, A. (2011). Governance of neuroscience: challenges and responses. In *Brain waves module 1: Neuroscience, society and policy* (pp. 87–96). London: Royal Society.

Tafolla, T. J., Trachtenberg, D. J., & Aho, J. A. (2012). From niche to necessity: Integrating nonlethal weapons into essential enabling capabilities. *JFQ: Joint Force Quarterly, 66*(3), 71–79.

Tennison, M., & Moreno, J. D. (2012). Neuroscience, ethics, and national security: The state of the art. *PLoS Biology, 10*(3), 1–4.

Trevan, T. (2012). Do not censor science in the name of biosecurity. *Nature, 486*, 299.

Tucker, J. B. (Ed.). (2012). *Innovation, dual use and security: Managing the risks of emerging biological and chemical technologies*. Cambridge, MA: The MIT Press.

United Kingdom. (2012). *The convergence of chemistry and biology: Implications of developments in neurosciences* (Working Paper No 1). Meeting of states parties to the convention on the prohibition of the development, production and stockpiling of bacteriological (biological) and toxin weapons and on their destruction. BWC/MSP/2012/MX/WP.1. Geneva: United Nations.

Vogel, K. M. (2008). Framing biosecurity: An alternative to the biotech revolution model? *Science and Public Policy, 35*(1), 45–54.

Wheelis, M., Rozsa, L., & Dando, M. R. (2006). *Deadly cultures: Biological weapons since 1945*. Cambridge, MA: Harvard University Press.

Young, L., Camprodon, J. A., Hauser, M., Alvero, P.-L., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgement. *PNAS, 107*(15), 6753–6758.

International Legal Restraints on Chemical and Biological Weapons

115

Catherine Jefferson

Contents

Abstract

Therapeutic advances in neuropharmacology and drug delivery could be exploited for the development of incapacitating biochemical weapons. This chapter examines the international legal restraints on chemical and biological weapons, with a particular focus on arms control law, human rights law, and international humanitarian law. It also examines the legal implications of other potential neuroweapons, such as neural-interface weapons systems, with a particular focus on the challenges posed to notions of criminal responsibility.

Introduction

Advances in neuroscience and understandings of the brain promise great benefits to society, particularly in the treatment of neurological impairment and psychiatric

C. Jefferson
Department of Social Science, Health & Medicine, King's College London, London, UK
e-mail: Catherine.Jefferson@kcl.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_140,

disease. However, as an unfortunate precedent attests, every major innovation in science and technology has been exploited for hostile purposes, and the life sciences have proven to be no exception (Meselson 2000; Kelle et al. 2012). While advances in neuropharmacology and drug delivery may offer improved therapeutic options, these same advances may also suggest ways in which human health can be impaired and could be exploited to do harm through the development of chemical and biological weapons (Royal Society 2012; ► Chap. 114, “Neuroscience Advances and Future Warfare” this section). This chapter examines the body of international law in place that prohibits these weapons, with a particular focus on arms control law, human rights law, and international humanitarian law. Finally, it examines some of the legal implications of other potential neuroweapons, with a particular emphasis on criminal responsibility.

Arms Control Law

Chemical and biological weapons (CBW) are weapons that are intended to do harm by affecting life processes, through either the toxicity of chemicals or the infectivity of pathogenic microorganisms. CBW are prohibited by two mutually reinforcing treaties that ban the use, development, and procurement of these weapons – the 1993 Chemical Weapons Convention (CWC) and the 1972 Biological Weapons Convention (BWC). As Table 115.1 illustrates, there is some overlap between the agents that are covered by these two treaties, and while there exist a number of distinguishing features between chemical weapons on the one hand and biological weapons on the other (such as toxicity versus contagiousness or incubation period), they are perhaps more appropriately conceptualized as a threat spectrum rather than distinct categories of weapons.

The norm against the use of poison and disease as a weapon can be traced throughout history and across various different cultures, but the first multilateral agreements to prohibit the use of poison or poisoned weapons in war were the Hague Conventions of 1899 and 1907, which also included provisions against the use of asphyxiating gas (Jefferson 2009). While the Hague Conventions were unsuccessful in averting the wide-scale use of gas in the First World War, the prohibition on CBW was reaffirmed with the signing of the Geneva Protocol of 1925, which remains in effect today and has provided the normative basis for the development of the CWC and BWC. Yet despite the legal regime that is in place to prohibit CBW, challenges remain. Of particular interest to this chapter is the perceived ambiguity within the provisions of the CWC, which, teamed with the perception that advances in neuroscience could offer feasible “incapacitating” biochemical weapons, risks undermining the regime against CBW.

Table 115. 1 CBW spectrum

Classical CW agents	Bioregulators Peptides	Toxins	Genetically Modified Organisms	Traditional BW agents
Chlorine Cyanide Phosgene Mustard Nerve Agents	Substance P Neurokinin A	Saxitoxin Ricin Botulinum Toxin	Modified/tailored bacteria or viruses	Bacteria Viruses Rickettsia Fungi
		Biological Weapons Convention		
	Chemical Weapons Convention			
	Poison			Infect

Adapted from Pearson (2002)

Chemical Weapons Convention

The 1993 CWC provides a comprehensive ban on chemical weapons, including a ban on their development, production, and stockpiling, as well as use. However, the CWC contains a number of exceptions whereby the production and use of toxic chemicals is permitted, being:

(a) Industrial, agricultural, research, medical, pharmaceutical or other peaceful purposes;
(b) Protective purposes, namely those purposes directly related to protection against toxic chemicals and to protection against chemical weapons;
(c) Military purposes not connected with the use of chemical weapons and not dependent on the use of the toxic properties of chemicals as a method of warfare;
(d) *Law enforcement including domestic riot control purposes*. (CWC Article 2.9)

It is through this law enforcement exemption that ambiguity has arisen. First, the CWC does not state explicitly what is meant by "law enforcement" or what law States may enforce, where or under what circumstances. As such, Article 2.9(d) could be interpreted to permit the use of toxic chemicals to enforce domestic law extra-jurisdictionally or to enforce international law (Meselson and Robinson 1994). However, the CWC is clear that riot control agents are prohibited as a method of war (CWC Article 1.5).

Secondly, and more presciently for this discussion, is the ambiguity in the interpretation of the range of toxic chemicals permissible for law enforcement purposes. While, as noted above, riot control agents are prohibited as a method of war, Article 2.9(d) protects the use of chemicals such as tear gas for domestic riot control purposes in the context of law enforcement. Riot control agents are defined

by the CWC as "any chemical not listed in a Schedule, which can produce rapidly in humans sensory irritation or disabling effects which disappear within a short time following termination of exposure." However, Article 2.9(d) stipulates that the production and use of toxic chemicals may be used for law enforcement *including* domestic riot control purposes, yet the Convention provides no definition of what these other chemicals may be. This ambiguity could thus be interpreted to suggest that the range of toxic chemicals permitted for law enforcement goes beyond riot control agents to also encompass so-called incapacitating chemical agents, i.e., centrally acting agents intended to cause rapid but more prolonged incapacitation through "loss of consciousness, sedation, hallucination, incoherence, paralysis, disorientation or other such effects" (Royal Society 2012).

The CWC distinguishes between three classes of toxic chemicals and their precursors. Schedule 1 chemicals have few or no uses outside of chemical weapons; Schedule 2 chemicals have legitimate small-scale applications; and Schedule 3 chemicals have large-scale uses apart from chemical weapons. The CWC is explicit that States Parties are prohibited from producing, acquiring, retaining, or using Schedule 1 chemicals for law enforcement purposes (CWC Verification Annex VI.2). Schedule 2 and 3 chemicals are permitted but are subject to Article 2.1(a) of the CWC, which restricts the development, possession, and use of toxic chemicals according to "types and quantities" consistent with permitted purposes. As Meselson and Robinson ask:

> Is the Convention really to be read as allowing any non-Schedule-1 toxic chemical or precursor to be developed, produced, weaponized, stockpiled or traded, so long as it is said to be for 'law enforcement purposes'?. (Meselson and Robinson 1994)

Experts in the field of CBW control and nonproliferation argue that such an interpretation is untenable and that only riot control agents should be considered permissible for law enforcement purposes, with the special exception of toxic chemicals used as a means of capital punishment (i.e., lethal injection). The special case of chemical executions in the USA, it is argued, should not be exploited for wider interpretations of law enforcement that go beyond riot control agents (Kelle 2012a).

On the other end of the spectrum, there are strong indications of State interest in the development of incapacitating chemical agents, ostensibly for law enforcement purposes (Crowley 2009). One of the more widely documented examples is the use by the Russian Federation of an incapacitating chemical agent reportedly a derivative of fentanyl, believed to be a mixture of carfentanil and remifentanil (Royal Society 2012). In October 2002, a group of armed Chechen separatists stormed the Dubrovka theater in Moscow and took approximately 830 people hostage, demanding withdrawal of Russian troops from Chechnya. On the third day of the hostage crisis, Russian special forces disseminated the incapacitating chemical agent through the theater's ventilation system, incapacitating both the hostages and the hostage-takers. The theater was then stormed by troops who killed the hostage-takers and freed the hostages. While the siege was brought to an end, 129 of the hostages died as a result of complications following the use of the incapacitating chemical agent and many others suffered long-term effects (Wax et al. 2003).

The perception of the need for improved operational capabilities, for example, in scenarios such as counterterrorism and counterinsurgency, teamed with insights gained from advances in neuroscience, is continuing to fuel State interest in incapacitating chemical agents. There are a wide range of neuropharmacological agents that could be considered as candidate agents for incapacitation, including opioids such as fentanyl and its derivatives, benzodiazepines, alpha-2-adrenoceptor agonists, and neuroleptic anesthetics. Advances in the knowledge of bioregulatory peptides could also offer novel means of incapacitation. While biochemical compounds that control vital homeostatic systems, such as temperature, sleep, and blood pressure, occur naturally at low concentrations, they can be toxic at higher concentrations. For example, research conducted on the effect of the aerosolized neuropeptide substance P found it to be fatally toxic when absorbed into the lungs of test animals (Koch et al. 1999). Developments in understandings of bioregulators and receptor systems could therefore be exploited to induce unconsciousness. Recent research into orexin receptor antagonists for the treatment of insomnia, for example, could suggest possible applications for the development of an incapacitating chemical agent (Royal Society 2012).

Despite these potential advances, numerous studies have highlighted the challenges of creating a genuinely "nonlethal" incapacitating chemical agent. One of the biggest challenges is the dose–response problem, that is, for an agent to be effective in its goal of reliable incapacitation, the dose required would generally be expected to cause a level of fatality. This was clearly demonstrated in the use of an incapacitating chemical agent by the Russian Federation during the theater siege. As one commentator has noted:

> The fentanyl, however, killed approximately 130 hostages – a fatality rate of 16 %, more than twice the fatality rate of "lethal" chemical weapons used on World War I battlefields. (Fidler 2005)

Determining an acceptable threshold of fatality is problematic, and even if an acceptable safety margin could be established under clinical control, it is doubtful that this could be replicated in an operational scenario. The dissemination of the agent in tactical situations would require pulmonary rather than intravenous delivery, which raises challenges for ensuring uniform dose delivery. Even if uniform delivery could be established, the target, being thus incapacitated, will remain in the contaminated area, possibly leading to prolonged exposure unless the person or the agent is removed (BMA 2007). As a recent UK Royal Society report summarizes:

> In addition to the technical challenge of combining an agent of sufficient safety margin with a dose controlled delivery system, uncontrollable variables such as variability of target population, secondary injury (eg, airway obstruction) and the need for medical aftercare make the feasibility of a totally safe incapacitating chemical agent unlikely. (Royal Society 2012)

Even if these technical challenges could be surmounted, when interpreting Article 2.9(d) of the CWC, States should take in account the object and purpose of the Convention: "…for the sake of all mankind, to exclude completely the

possibility of the use of chemical weapons...” (CWC Preamble). As various commentators have noted, the proliferation of incapacitating chemical weapons could weaken restraints on the use of lethal forms of chemical weapons, thereby threatening to undermine the regime against CBW and the norm against the use of poison and disease as a weapon:

> The continued development and use of toxic chemicals as weapons for law enforcement is likely to present broad and unpredictable risks for security, including inevitable proliferation... Depending on the extent of proliferation there could be the risk of an “arms race” of new chemical weapons and defensive countermeasures, which could be accentuated by any military acquisition of these weapons. (ICRC 2012)

At the time of writing, the Third Review Conference of the CWC is approaching (April 2013), and it is hoped that action will be taken toward a common understanding of the law enforcement provision and the status of incapacitating chemical agents.

Biological Weapons Convention

The 1972 BWC provides a comprehensive ban on biological and toxin methods of warfare. The BWC prohibits “microbial or other biological agents, or toxins whatever their origin or method of production, that have no justification for prophylactic, protective or other peaceful purposes” (BWC Article 1.1). The BWC is therefore comprehensive in its scope, and since certain candidate incapacitating agents, such as bioregulators and peptides, would be considered biological agents or toxins (see Table 115.1), these would also be prohibited by the BWC.

One potential area of ambiguity within the BWC is the definition of “peaceful purposes” and “hostile purposes.” It is unclear, for example, how the use of incapacitating biochemical agents for counterterrorism, counterinsurgency, or operations other than war would be regulated by the BWC. However, a number of States Parties to the BWC have highlighted the dangers to the BWC of the misuse of bioregulators and peptides in the development of incapacitating agents.

Human Rights Law

Given the ambiguities within the CWC regarding the potential development of incapacitating chemical agents, some commentators such as the ICRC have looked to other bodies of law to reinforce limitations on the use of toxic chemicals in the context of law enforcement.

International human rights law prohibits torture and cruel, inhuman, or degrading treatment or punishment and protects the right to life by placing constraints on the use of force. While it is not totally prohibited for an authorized agent of the State to have recourse to lethal force, the use of any weapon in a law enforcement context must be based on the criteria of lawfulness, necessity, and

proportionality (Casey-Maslen 2011). Given the risks of lethality, the use of incapacitating chemical agents for law enforcement purposes could therefore be called into question under international human rights law.

The death of 129 hostages following the use of an incapacitating chemical agent by the Russian special forces in an attempt to resolve the 2002 Moscow theater siege, and the failure of the authorities to disclose the chemical used, has been subject to scrutiny under international human rights law. In July 2003, eighty former hostages filed a complaint to the European Court of Human Rights, claiming their right to life had been violated by the actions of the Russian authorities.

In 2011 the European Court of Human Rights ruled that the Russian government had violated the right to life of the hostages through inadequate planning and implementation of the rescue operation but not on the basis of the chemical agent used. As Kelle has argued, this raises a number of problems:

> Apart from the fact that the use of a toxic chemical by Russian Special Forces clearly did not leave any chance of survival to 125 of the hostages, such an effects-based understanding of indiscriminate use of force seems strangely at odds with the established principles of the law of armed conflict. (Kelle 2012b)

A further problem relates to the Court's distinction between lethal and nonlethal use of force based on intention, i.e., it is assumed that the gas was a nonlethal incapacitating weapon since it "was probably not intended to kill the terrorists or hostages" (ECHR 2011). On the other hand, the Court found that the gas "remained a primary cause of death of a large number of victims" (ECHR 2011).

Some commentators have argued that this ruling can be interpreted as legitimizing the use of incapacitating chemical agents for law enforcement (Kelle 2012b). On the other hand, it reinforces the difficulties of actual implementation, as the ICRC note:

> . . .it is evident that the dose of a chemical delivered cannot be controlled in such a tactical situation and that it is extremely difficult, if not impossible, in such situations to provide the immediate medical care that might be characterised as adequate to protect life. (ICRC 2012)

International Drugs Control Law

International drugs control treaties are another source of international law that limits States' use of toxic chemicals. The 1961 Single Convention on Narcotic Drugs, as amended by the 1972 Protocol, the 1971 Convention on Psychotropic Substances, and the 1988 UN Convention Against the Illicit Traffic in Narcotic Drugs and Psychotropic Substances place strict limits on certain controlled toxic chemicals. These treaties limit the possession, use, trade, and production of narcotic drugs exclusively to medical and scientific purposes. The legitimacy of States Parties developing and using narcotic drugs for activities such as law enforcement is therefore called into question (ICRC 2012).

International Humanitarian Law

In addition to restraints on the potential law enforcement applications of neuroscience, potential military applications of neuroscience are also constrained by international humanitarian law (IHL), sometimes referred to as the laws of war or law of armed conflict, which seeks to limit the effects of armed conflict. IHL is primarily comprised of the Geneva Conventions of 1949 and the Additional Protocols of 1977 relating to the protection of victims of armed conflicts as well as other agreements prohibiting or regulating the use of specific weapons, such as the CWC and BWC.

The Geneva Conventions of 1949 and the Additional Protocols of 1977 (Protocols I and II) provide an extensive regime for the protection of victims of armed conflict. Within this regime, a number of provisions are particularly relevant in considering the use of incapacitating chemical agents in the context of armed conflict. For example, Article 41 of Additional Protocol 1 and Common Article 3 of the Geneva Conventions require the humane treatment of all persons in enemy hands, including civilians and members of the armed forces placed *hors de combat* (i.e., incapable of hostile action) by sickness, wounds, detention, or any other cause. As Article 41 of Additional Protocol 1 stipulates:

> 1. A person who is recognized or who, in the circumstances, should be recognized to be hors de combat shall not be made the object of attack.
> 2. A person is hors de combat if:
>   (a) he is in the power of an adverse Party;
>   (b) he clearly expresses an intention to surrender; or
>   (c) *he has been rendered unconscious or is otherwise incapacitated by wounds or sickness, and therefore is incapable of defending himself*;
>
> Provided that in any of these cases he abstains from any hostile act and does not attempt to escape. (Protocol 1, emphasis added)

The use of incapacitating chemical agents that render members of the armed forces *hors de combat* would therefore mean that affected persons must not be attacked by any other method from that time on. This presents particular concerns given that military personnel may not be able to recognize the signs of incapacitation. As an ICRC study asks:

> Due to the lack of physical evidence of injury, would an incapacitated person be more easily subjected to continued attack than persons rendered *hors de combat* by conventional weapons? (ICRC 2010)

The use of incapacitating chemical agents in armed conflict is also constrained by the principle of distinction:

> Rule 11: Indiscriminate attacks are prohibited.
> Rule 12: Indiscriminate attacks are those:
> (a) Which are not directed at a specific military objective;
> (b) Which employ a method or means of combat which cannot be directed at a specific military objective; or
> (c) Which employ a method or means of combat the effects of which cannot be limited as required by international humanitarian law;
>
> And consequently, in each such case, *are of a nature to strike military objectives and civilians or civilian objects without distinction*. (ICRC 2005, emphasis added)

Thus belligerents within a conflict must at all times distinguish between civilians and combatants, and attacks may only be directed against combatants. The use of incapacitating chemical agents in mixed combatant/civilian scenarios would therefore breech this principle (Herby 2007).

Other Neuroweapons?

The weaponization of neuroscience is not limited to the potential development of new chemical and biological weapons. One area of neuroscience that is attracting considerable attention is the potential for militaries to exploit neurotechnologies such as brain-machine interfaces that enable the direct neurological control of weapons systems. However, while neurologically controlled weapons are not necessarily prohibited by international humanitarian law per se, they do raise significant questions over notions of criminal responsibility.

Advances in brain-machine interfaces are enabling brain signals to be connected to specific hardware or software systems. It is conceivable that brain-machine interfaces could be utilized to enable direct neurological control of remotely operated robots, unmanned vehicles, or weapons systems. While legal responsibility for a military activity remains with the last person to issue the command, concerns have been expressed that advances in neuroscience could enable the development of weapons systems that blur the distinction between thought and action (White 2008). For example, electroencephalogram (EEG) studies have revealed that image processing can occur much more rapidly than the subject is consciously aware of, which could have implications for target detection (Royal Society 2012). A neurally interfaced weapons system that responds to neural markers could therefore provide advantages in terms of speed and accuracy of target detection but would remove the element of conscious human deliberation.

State practice has established that a serious violation of international humanitarian law in both international and non-international armed conflicts constitutes a war crime (ICRC 2005). War criminals may be brought before national courts, international criminal tribunals, or the permanent International Criminal Court, which is governed by the Rome Statute. However, if a war crime is committed by a semiautonomous weapons system, then determining criminal responsibility becomes more problematic. For example, the war crime of wilful killing could be committed if a neurally interfaced weapons system impacts on the ability to distinguish between legitimate combatants and civilians or people who have surrendered and thereby gained protection under international humanitarian law. Liability for such a war crime is seemingly unclear. As White has argued:

> If neural-interfaced weapons were designed to fire at the time of recognition rather than after the disambiguation process, a process that would likely need to occur for the pilot to differentiate between combatants and protected persons, the pilot firing them presumably would lack criminal accountability for the act implicit in wilful killing. *Because of the way brain-interfaced weapons may interrupt the biology of consciousness, reasonable doubt*

> *may exist as to whether an actor performed a conscious act in the event of a contested incident*. (White 2008, emphasis added)

This ambiguity has led to calls for an expanded doctrine of command responsibility in which designers and engineers of weapons systems could also be held responsible for aiding and abetting a war crime (White 2008). However, typically the responsibilities of designers are discharged once the system is certified by the relevant national authority (Gillespie and West 2010). This assumes that a system's basic principles of operation have been shown to be lawful as part of its release into service and that the individual giving orders will ensure its continued lawful use (MOD 2011). Legal responsibility will thus remain with the last individual to issue the command. However, given the complexities of some of the conceivable applications of neurally interfaced weapons systems, it is questionable whether such systems will continue to behave in a predictable manner after the command has been issued.

Neurally interfaced weapons systems therefore pose a number of legal dilemmas. Some commentators have argued for an outright prohibition on semi- and fully autonomous weapons (such as, the International Committee for Robot Arms Control). However, such a prohibition could hinder the development of weapons systems that may prove to be more accurate than existing capabilities. A balanced approach to future policy must be maintained in which parameters for accountability and acceptable machine control is established.

Conclusion

Advances in neuroscience offer a range of significant benefits to society. However, many of these developments could also be exploited for hostile applications, such as in the development of incapacitating chemical weapons. This chapter has provided an overview of the legal restraints on these weapons in both military and law enforcement contexts, drawing on arms control law, human rights law, and international humanitarian law. It has also examined some of the legal implications of potential developments in neurally interfaced weapons systems.

Despite the existence of a comprehensive regime of restraints against the development and use of incapacitating chemical weapons, studies have revealed a general lack of awareness among neuroscientists of the CWC and BWC. It has been noted by a number of commentators – and is formally recognized by the implementing bodies of the two treaties – that awareness raising among scientists is crucial for the effective implementation of the CWC and BWC. Greater awareness of the potential malign applications of neuroscience research and the legal regime that such applications would be subject to should therefore be encouraged.

Cross-References

▶ Neuroscience Advances and Future Warfare

References

BMA. (2007). *The use of drugs as weapons: The concerns and responsibilities of healthcare professionals*. London: British Medical Association.

Casey-Maslen, S. (2011). *Weapons termed 'non-lethal'; and international human rights law*. Geneva: Geneva Academy of International Humanitarian Law and Human Rights.

Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (Biological Weapons Convention), 10 April 1972.

Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction (Chemical Weapons Convention), 13 January 1993.

Crowley, M. (2009). *Dangerous ambiguities: Regulation of riot control agents and incapacitants under the chemical weapons convention*. Bradford: Bradford Non-Lethal Weapons Research Project.

ECHR. (2011). *Use of gas against terrorists during Moscow theatre siege was justified, but the rescue operation afterwards was poorly planned and implemented. Press release issued by the Registrar of the Court*. Strasbourg: European Court of Human Rights.

Fidler, D. P. (2005). The meaning of Moscow: 'Non lethal' weapons and international law in the early 21st century. *International Review of the Red Cross, 87*, 525–552.

Gillespie, T., & West, R. (2010). Requirement for autonomous unmanned air systems set by legal issues. *The International C2 Journal, 4*, 1–32.

Herby, P. (2007). Protecting and reinforcing humanitarian norms: The way forward. In A. M. Pearson, M. I. Chevrier, & M. Wheelis (Eds.), *Incapacitating biochemical weapons: Promise or peril?* Lanham: Lexington.

ICRC. (2005). *Customary international humanitarian law, Volume 1: Rules*. Cambridge: Cambridge University Press.

ICRC. (2010). *Incapacitating chemical agents: Implications for international law*. Geneva: International Committee of the Red Cross.

ICRC. (2012). *Toxic chemicals as weapons for law enforcement: A threat to life and international law?* Geneva: International Committee of the Red Cross.

Jefferson, C. (2009). *The taboo of chemical and biological weapons: Nature, norms and international law* (DPhil Dissertation). University of Sussex.

Kelle, A. (2012a). Legally incapacitated, politically outmanoeuvred. *Bulletin of the Atomic Scientists*. http://www.thebulletin.org/web-edition/columnists/alexander-kelle/legally-incapacitated-politically-outmaneuvered. Accessed 30 Nov 2012.

Kelle, A. (2012b). The message from Strasbourg. *Bulletin of the Atomic Scientists*. http://www.thebulletin.org/web-edition/columnists/alexander-kelle/the-message-strasbourg. Accessed 30 Nov 2012.

Kelle, A., Nixdorff, K., & Dando, M. (2012). *Preventing a biochemical arms race*. Stanford: Stanford University Press.

Koch, B. L., Edvinsson, A. A., & And Koskinen, L. O. (1999). Inhalation of substance P and thiorphan: Acute toxicity and effects on respiration in conscious guinea pigs. *Journal of Applied Toxicology, 19*, 19–23.

Meselson, M. (2000). Averting the hostile exploitation of biotechnology. *The CBW Conventions Bulletin, 48*, 16–19.

Meselson, M., & Robinson, J. P. (1994). New technologies and the loophole in the convention. *Chemical Weapons Convention Bulletin, 23*, 1–2.

MOD. (2011). *Joint Doctrine Note 2/11: The UK approach to unmanned aircraft systems*. Shrivenham: Ministry of Defence.

Pearson, G. (2002). *Relevant scientific and technological developments for the first CWC review conference: The BTWC review conference experience*. CWC review conference paper No. 1. University of Bradford: Department of Peace Studies.

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.
The Royal Society. (2012). *Brain waves 3: Neuroscience conflict and security*. London: The Royal Society.
Wax, P. M., Becker, C. E., & Curry, S. C. (2003). Unexpected 'gas' casualties in Moscow: A medical toxicology perspective. *Annals of Emergency Medicine, 41*, 700–705.
White, S. E. (2008). Brave new world: Neurowarfare and the limits of international humanitarian law. *Cornell International Law Journal, 41*, 177–210.

Biosecurity as a Normative Challenge 116

Tatyana Novossiolova

Contents

Abstract

The ongoing progress of biotechnology constitutes a major governance challenge, not least because the same advances that promise to enhance human welfare could potentially enable the development of novel biological weaponry systems. To address the multifaceted security concerns arising from the life sciences, a range of top-down policy initiatives and legally binding regulations have been introduced, but their overall impact on the practices of life scientists has remained limited. Given that laws in every sphere of activity are dependent upon the vitality of social and/or professional norms, the chapter aims to enquire into the normative foundation of the biosecurity regulations. It contends that there is a disjuncture between the conduct of life scientists and the rules pertaining to biosecurity stemming from the lack of corresponding norms in the professional culture of life science research. This disjuncture largely manifests itself in three forms, namely, ignorance of the existing biosecurity regulatory framework, arrogance motivated mainly but not

T. Novossiolova
Bradford Disarmament Research Centre, Division of Peace Studies, University of Bradford, Bradford, UK
e-mail: t.a.novossiolova@bradford.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_142,

exclusively by the belief that science should enjoy unconstrained freedom, and acts of open defiance of the rules. The chapter concludes by examining the value of biosecurity education and awareness raising in addressing the security challenges posed by biotechnology and fostering a culture of research that is keenly aware of and responsive to the norm of biological nonproliferation.

Introduction

The advancement of biotechnology over the past few decades promises to bring tremendous benefits by responding to health, societal, and environmental ills. At the same time, however, this progress poses fundamental ethical, legal, and social challenges, not least because the very same advances that could enhance public welfare could potentially facilitate the development of novel biological weapons and/or be misused for bioterrorist purposes and thus pave the way for the wholesale militarization of the life sciences. In February 2012, the British Royal Society published Brain Waves Module 3: Neuroscience, Conflict and Security, a report outlining the potential military and law enforcement applications of novel discoveries in the field of neuroscience. According to the report, such applications broadly pursue two main objectives, namely, improving the efficiency of one's own forces (performance enhancement) and diminishing the performance of one's enemy (performance degradation) (The Royal Society 2012). It further suggested that there are at least several major developments, including neuropharmacology, functional neuroimaging, and neural interference systems that immediately raise important ethical and legal issues (The Royal Society 2012). Unfortunately, to date, engagement among neuroscientists with the ethical, social, and legal aspects of their work has been limited. Far from being an exception, this trend appears common among life scientists in general. A series of interactive seminars carried out in 16 different countries with a several thousands of life scientists in over 110 different departments revealed that there was little evidence that the participants:

1. Regarded bioterrorism or bioweapons as a substantial threat
2. Considered that developments in life sciences research contributed to biothreats
3. Were aware of the current debates and concerns about dual-use research
4. Were familiar with the Biological and Toxin Weapons Convention (BTWC) (Dando and Rappert 2005; Rappert 2007, 2009)

By contrast, following the 9/11 assaults, the 2001 "anthrax letters" attacks and most recently the controversy surrounding the production of mammalian-transmissible avian influenza (H5N1), attempts to address the security implications of novel biotechnology advances have been high on the political agenda. Besides the 1975 BTWC, which outlaws a whole class of weapons of mass destruction (WMD), a whole range of legally binding instruments, regulations, codes, and guidelines have been devised at international, national, and institutional level in order to improve the governance of dual-use science and technology and ensure that the life sciences are utilized only for "peaceful, protective, and prophylactic purposes." While those documents have formally expanded the scope of

responsibilities incumbent upon those engaged in the life sciences, their practical implementation has been slow and patchy. A major hurdle in the process has been a widely shared perception among life scientists that biosecurity regulations are largely burdensome and unnecessary, creating extra paperwork and threatening to stifle innovation.

The purpose of this chapter is to account for the existing disjuncture between the practices and attitudes of life scientists, on the one hand, and biosecurity regulations, on the other. It advances the argument that the disjuncture stems from the fact that the regulations are not underpinned by corresponding norms in the professional culture of life scientists, which in turn impacts on their integrity and effectiveness. Section "The Power of Norms" elucidates the interplay between laws and norms, demonstrating the ways in which they reinforce one another. Section "Biosecurity and Risk in Science Practice" then conceptualizes biosecurity and examines the scientists' perceptions of risks of vis-à-vis the biosecurity discourse. Section "Biosecurity Regulations: Role and Limitations" gives an overview of biosecurity regulations, their role, and limitations. Section "The Disjuncture Manifested" looks into some the practices of life scientists and how they diverge from the established regulatory framework. The chapter concludes that addressing the security challenges posed by biotechnology requires a culture of research keenly aware of and responsive to the norm of biological nonproliferation.

The Power of Norms

Vickers suggests that there are three types of constraints that shape human motivation and action: logistical constraints that define the limits to what one can do, taking into account possible risks and uncertainty; constraints imposed by the wishes and expectations of others, including those imposed by powers of authority that define what one must or must not do within a society, group, or an organization; and constraints created by one's own personality system that define what one ought to do in order to meet one's own aspirations and expectations of oneself (Vickers 1972). In other words, what one can and cannot do, must and must not do, and ought and ought not do are subject to constraints imposed by circumstances, by other people and oneself, respectively (Vickers 1972). Thus, physical capacity and social status constitute logistical constraints, laws and norms serve as regulatory constraints, and morals are among the personal constraints that guide individual's deeds. This section will focus exclusively on the second type of constraints which in the widest sense relate to social obligation.

Laws are fundamental ingredients for the functioning of any society, not least because they draw a clear distinction between what must and must not be done and, as such, constitute the formal rules of behavior that every individual must obey. By design, the primary role of laws is to provide a basis for social order, which is why the extent of their effectiveness is directly proportionate to the degree to which they are abided by. In order to ensure maximum compliance with the law and limit the instances of disobedience, states rely upon various

enforcement mechanisms, including the use of police force and imposition of penalties. Yet, the value of legal rules notwithstanding, it is naïve to assume that law is the only or indeed the most important source of order within a social grouping. This is so, for people do not generally refrain from certain activities, such as stealing or committing assault, solely on the grounds of fear from incurring sanctions afterwards but rather because it is in their interest to do so. In other words, by accepting certain constraints on their personal freedom, individuals signal their inclination to cooperate with other members of society and contribute to the common good, thus enjoying benefits which otherwise will be unattainable. Not only does free riding then become a costly option but it also loses its attractiveness, as "cheaters" are hardly tolerated and likely to be ostracized (Becker 1996). A powerful social stigma on certain actions will undoubtedly have direct implications for the establishment of social order. To be sure, the members of a society in which theft and murder are commonly viewed as unacceptable acts that merit denunciation are much more likely to enjoy a relative degree of safety than those living in a society utterly dependent on law enforcement for tackling such problems. The shared expectation that cooperative behavior is rewarded, whereas cheating is penalized, facilitates the emergence of informal and unspoken rules which shape social relations and regulate the conduct of affairs in both public and private sphere – norms.

Norms are concrete, specific, and tacit standards (Vickers 1973) that signify what a particular social group considers acceptable and what unacceptable modes of behavior. Unlike laws, they are not dependent on government for either promulgation or enforcement (Posner and Rasmusen 1999). Rather, norms arise from and crystallize in the emergence of gradual consensus (Posner and Rasmusen 1999) as to what the members of a given group perceive as "normal" and are sustained by a shared recognition of their importance. Norms carry significant regulatory weight, insofar as they reflect the ethos of the group, that is, the beliefs, values, and morals shared by its members, and instances of noncompliance may have severe consequences. Thus, conformity with established norms both plays a paramount role in the maintenance of social order and constitutes an indispensable precondition for the functioning of law. As Tamanaha (2001) has pointed out, the "state legal system would not even exist, were it not for an already stable and effective baseline provided by the unarticulated substrate, shared norms, instrumental behaviour and consent."

Norms are inevitably subject to cultural and historical contingencies and tend to vary across communities. Far from being static, they evolve and change in parallel with social processes and practices. For their part, laws by and large, follow the dynamics of social norms and adapt accordingly. Laws that guarantee women's rights have only emerged as a result of the persistent efforts of the emancipation movements over the past century which challenged the prevalent norms of male superiority. Similarly, the collapse of the apartheid regime in South Africa would hardly have occurred had not it been for the lack of popular support for the established order (Whitman 2009).

The reverse process, however, whereby the creation of legal rules precedes or seeks to foster the development of social norms, is not so common. The history of

the human rights regime is a case in point. Brought about by the struggle of individuals everywhere to delegitimize the relations of absolute power, the human rights regime has played a major role in the abolition of colonialism, total war, and slavery. At the same time, the fact that practices as gruesome as female genital mutilation (FGM) and honor killing still persist in certain cultures even after the provision of the 1948 Universal Declaration of Human Rights has been formally codified in international and regional treaties and national legislation is indicative of the complexities of altering social norms. The obstacles to effective norm entrepreneurship are further illustrated in the reluctance of democratic governments to impose legal rules which are inconsistent with the prevalent social norms. Hence, the decision of President Lyndon Johnson to sign the Civil Rights Act of 1964 is often cited as an example of political courage, not least because of the public outrage that followed the new law, especially in the southern states. Needless to say, the Act required a substantial degree of enforcement in order to have an effect on lessening racial discrimination and segregation.

As the discussion above indicates, laws and norms are mutually reinforcing. It is also clear that while legal rules largely evolve from established customary practices, fostering norms via formal codes and regulations is a slow and arduous endeavor which often needs to overcome considerable resistance and, as such, can hardly succeed only by dint of enforcement. In light of those revelations, the subsequent sections aim to inquire into the normative foundation of the biosecurity regulations designed to address the governance challenges arising from the enormous dual-use potential of modern biotechnology.

Biosecurity and Risk in Science Practice

According to Dando (2010), the specter of threats to human health can be divided into three broad categories: naturally occurring disease, disease resulting from the accidental release of pathogens, and deliberately caused disease. Each of those threats is addressed through a distinct approach utilizing public health, laboratory biosafety, or biosecurity measures, respectively. Based on this understanding, biosecurity is closely linked to the idea of the "web of preventive policies" centered on the prohibition of the malevolent misuse of the life sciences embodied in the Biological and Toxin Weapons Convention (BTWC) (Novossiolova and Whitby 2011). As such, it is the objective of a whole range of rules and measures, including export controls, biodefense, and national implementation of the BTWC, designed to prevent the hostile release of pathogens by either state or non-state actors.

In light of the enormous dual-use potential of modern biotechnology and the resultant security challenges that strike right at the heart of the biological weapons nonproliferation regime, over the past decade novel biosecurity regulations which substantially expand the scope of responsibilities incumbent on those engaged in the life sciences have been introduced. The 1975 Biological and Toxin Weapons Convention (BTWC) is considered the cornerstone of the biological nonproliferation regime. Article I of the Convention formally prohibits work on pathogens and

toxins that have no justification for "prophylactic, protective, or other peaceful purposes." Yet to the rapid advancement of biotechnology poses a considerable threat to the regime, not least because it allows for benignly intended life science research to be potentially exploited for hostile purposes, including the development of biological weapons and bioterrorism.

Following the events of 9/11 and the "anthrax letters" attacks several states, most notably the USA, have developed and implemented strict rules and policies, designed to prevent the hostile exploitation of benignly intended scientific research. Two high-level reviews – Biotechnology Research in an Age of Terrorism (The Fink Report) and *Globalization, Biosecurity, and the Future of the Life Sciences* (The Lemon-Relman Report) – sought to draw attention to the issues of dual use by defining several potentially dangerous classes of research that warranted institutional review and oversight prior to being performed, and highlighting the important role that practicing life scientists could play in improving the governance of biotechnology.

Yet to date, there has been limited engagement of the life science community with biosecurity issues. Indeed, it is no exaggeration to say that while dual-use research may be a significant matter for security specialists and politicians, it is hardly a source of moral dilemmas for practicing researchers, as both the Fink and the Lemon-Relman Committees have suggested (NRC 2004, 2006). On the contrary, biosecurity concerns arising from cutting-edge biotechnology remain broadly unacknowledged and sometimes tend to be even purposefully ignored by those on the front lines of research. The recent controversy generated by two studies describing the creation of mammalian-transmissible avian influenza (H5N1) is indicative of both the propensity of life scientists to overstate the benefits of scientific research at the expense of potential security risks and the prevalent lack of awareness of biosecurity issues among the life science community. Had the vocal advocates of research on contagious H5N1 virus read the BTWC, they would have known that work involving the aerosolization of pathogens has direct implications for Article III of the Convention which states that State Parties undertake:

> not to transfer to any recipient whatsoever, directly or indirectly, and not in any way to **assist**, encourage, or induce any State, group of States or international organisations to manufacture or otherwise acquire any of the agents, toxins, weapons, equipment or means of delivery specified in Article I of the Convention. [emphasis added]

Curiously, one of the lead researches on the team in the Netherlands, where the more dangerous study was performed, was part of the group that prepared the national Code of Conduct for Biosecurity, which has the BTWC as its first appendix.

Unlike the instances of scientific misconduct, such as falsification, fabrication, and plagiarism, which are usually met with condemnation within the research community, the violation of biosecurity regulations rarely causes outrage. An illustration of this trend is the case of Dr Thomas Butler and the massive support from scientists it attracted. Butler was charged with offenses under the US PATRIOT Act for mailing and transporting vials of plague on passenger aircraft

and by road without obtaining the necessary documents and was sentenced to two years in prison and thousands of dollars in fines, as a result (Enserink and Malakoff 2003). Nevertheless, the Federation of American Scientists launched a large-scale campaign urging researchers to sign a petition in Butler's defense and formally blamed the US government for scapegoating a prominent plague scientist (FAS 2006).

The reasons for the apparent disjuncture between the practices of life scientists and the biosecurity regulations are complex and, as such, can hardly be reduced to incapacity of researchers to apprehend potential hazards and arrive at moral judgment. To grasp the intricacy of the challenge posed by dual-use research, therefore, it is essential to move beyond ethical reasoning at individual level and consider the structural and cultural contingencies that shape the science enterprise and determine the perceptions of risk of those involved.

The issue of dual use is problematic on at least three levels. First, science research is inherently of dual-use character, since virtually any knowledge or piece of technology, however basic, can be applied for multiple purposes. Hence, the potential threat posed by dual-use research is not one of *nature* but of *kind*, for it is the scale of the peril and the gravity of the consequences that ring the security bell. At the next level is the question of intent. By and large, scientists seek to generate knowledge and products that will enhance human health and welfare. Given their genuine desire to serve the public, it is often beyond the perception of the average researcher that the results of their work can be of use for those who want to cause harm. At the third and final level, the challenge is informational one. The majority of scientists are hardly aware of the wider social and legal implications of their research, not least because they work in conditions (e.g., academia, industry) and for purposes (e.g., vaccine development) which place their perceptions of the risks of their activity primarily within the sphere of laboratory biosafety (e.g., the prevention of accidental release of microbes) but not further. By contrast, in their greatest part, the biosecurity regulations reflect the views and risk perceptions of lawmakers and defense and intelligence specialists, people traditionally preoccupied with guaranteeing national security and barely aware of the realities of life science research.

Biosecurity Regulations: Role and Limitations

At an international level the biosecurity regulatory architecture comprises three legally binding instruments: the UN Security Council Resolution 1540, which imposes obligations on all state to enact legislation to prevent the proliferation of weapons of mass destruction, including biological weapons, and their means of delivery (UNSC 2004); the International Health Regulations which seek to "prevent, protect against, control and provide a public health response to the international spread of disease in ways that are commensurate with and restricted to public health risks" (WHO 2005); and the BTWC. States are obliged to develop and enforce legal mechanisms accordingly and report their annual progress on the

national implementation of the documents provisions. At a national level some states have already developed comprehensive governance frameworks designed to address the challenges arising from dual-use life science research. In the USA, the US PATRIOT Act, the 2002 Public Health Security and Bioterrorism Preparedness and Response Act, and the 2008 Select Agents Regulations, for instance, complement one another in mitigating the risks of a possible bioterrorist attacks by regulating the access to, transport, and handling of potentially dangerous pathogens and toxins. Likewise, in Canada, the 2009 Human Pathogens and Toxins Act serves as the cornerstone of a "safety and security regime" established to prevent the malevolent misuse of pathogenic microbes. Other measures implemented for the purpose of obstructing the deliberate spread of disease include personnel reliability programs, background security checks, university vetting schemes (e.g., UK), and national biosecurity codes of conduct (the Netherlands). Most recently, the US government has introduced a new policy for oversight of dual-use research of concern, and it is expected that other countries will mirror this example (Editorial 2012a).

By design, the biosecurity regulations seek to ensure that benignly intended life science research is not misused for hostile purposes, including the development of biological weapons and bioterrorism. As such, the regulations serve a twofold function, insofar as they are intended both to promote scientific innovation and reinforce the norm of biological nonproliferation. Unlike biosafety rules that define acceptable laboratory practices and procedures, the biosecurity measures address the possible social and security implications of research "beyond the laboratory door." In other words, whereas the former provide guidelines as to how research should be conducted in practical terms, the latter tackle a broader scope of issues, including what kind of experiments should be done, who should have access to hazardous materials and agents, and how the research results should be disseminated. The utility of laboratory biosafety regulations notwithstanding, their role in the governance of dual-use research is limited, as several controversial experiments have demonstrated. It suffices to note that the creation of a highly virulent strain of Mousepox, the artificial synthesis of the polio virus, the recreation of the 1918 Spanish Influenza, and the production of both strains of contagious H5N1 virus have been conducted under appropriate laboratory conditions and in compliance with the established biosafety rules. Preserving the integrity of the biological weapons nonproliferation regime therefore requires that a wider span of responsibilities should be assigned to life scientists in order to guarantee that the results of their work are not misused to cause harm.

The Disjuncture Manifested

As already discussed, laws in every sphere of human activity are largely dependent on the vitality and plasticity of both social and professional norms. The case of the biosecurity regulations is particularly curious, not least because there is little evidence that a corresponding norm of biosecurity practice exists. In contrast to

laboratory biosafety rules that form an essential part of the training of life scientists and are now deeply embedded in the conduct of research, biosecurity regulations are seen as an unnecessary bureaucratic burden devised to impinge on the development of science. Among the criticisms leveled at the regulations by scientists is that the former are too narrow in scope and, as such, inevitably neglect important aspects of the culture of biotechnology research, including the practices of peer review and dissemination of results, both of which take place in a highly globalized environment. Those and other limitations of the existing regulatory mechanisms have been made explicit during the lengthy debate on the publication H5N1 studies, leading the editorial board of a prominent scientific journal to conclude that restricting the spread of sensitive dual-use research via censorship and data-sharing on a need-to-know basis is impracticable (Editorial 2012b).

But the implications of the H5N1 controversy are well more far reaching, not least because the intensive deliberations on the value of the experiments have exposed the negligible role that the biosecurity regulations play in the risk-benefit calculations of those engaged in the life sciences. This disjuncture generally manifests itself in three ways. Ignorance about the legal and social implications of dual-use research is by far one of the most common manifestations. As Working Paper No.20 submitted by Australia et al. to the Seventh Review Conference of the BTWC in December 2011 noted:

> Continued academic research on bioethics and awareness of biosecurity risks seem to confirm a generally limited level of awareness among life scientists in numerous institutions in numerous countries. Analysis of the reasons for this lack of awareness include, *inter alia*, the lack of university courses covering aspects related to the BWC and related (bio-) security issues, either because the curriculum developers do not consider the topic to be important or have difficulty fitting teaching material on biosecurity into what they claim is an already overcrowded curriculum, or because of a lack of expertise and access to relevant teaching material. (WP.20/BWC/2011)

Even though the paper emphasized that the pervasive lack of awareness of aspects related to biosecurity and the obligations of the Convention among life scientists has to be addressed more urgently, strategically, and comprehensively (WP.20/BWC/2011), little has been done so far to fill the education gap and sensitize scientists to the security challenges arising from dual-use technologies.

The disjuncture is further signaled by dint of arrogance. Given that the ultimate goal of science is the pursuit of truth which in turn assigns it a special status, some have argued that any attempt to curtail research is detrimental to human progress:

> We do know that the potential is there, but it is not through fear that we will stop H5N1 from becoming pandemic. The pursuit of knowledge is what has made humans resilient – a species capable of overcoming our worst fears. (Perez 2012)

As a result, ethical and social issues barely feature in the context of scientific debates and security risks tend to be downplayed as exaggerated and hypothetical. On rare occasions, arrogance may escalate into defiance of rules. For instance, it was quite telling that one of the lead scientists on the team in the Netherlands that created a contagious strain of the H5N1 virus publicly declared his readiness to

disregard the export control ban imposed on his study by the Dutch government and submit his manuscript for publication without obtaining the required license (Butler 2012).

Conclusion: Toward Education

Addressing the challenges posed by the rapid advancement and proliferation of dual-use science and technology requires the development of a culture of responsibility among those engaged in the life sciences. Still, it is at best myopic and at worst dangerous to assume that biosecurity regulations and oversight mechanisms per se can be effective in this endeavor given the lack of established corresponding norms of research practice. The process of fostering biosecurity norms in biotechnology is multidimensional, but it requires, at the very least, that those on the front lines of research are sensitized and responsive to the ethical, social, and legal implications of their work. Indeed, a growing body of evidence suggests that when made aware of the potential risks and dangers associated with their research, life scientists are serious about their broader responsibilities and eager to contribute their expertise to strengthening the biological weapon nonproliferation regime (Whitby 2012). Biosecurity education is thus one possible way of overcoming the existing normative deficit in the life sciences and facilitating the dialogue between researchers, the security community, and policymakers. It is worth mentioning that the value of security education for those working in sensitive fields of science, such as nuclear science, has long been recognized and tremendous progress toward creating a culture of responsibility has been made.

According to the International Atomic Energy Agency (IAEA 2008), fostering nuclear security culture is essential for ensuring that “the implementation of nuclear security measures receives the attention warranted by their significance.” In order to help member states develop and implement relevant policies that maximize nuclear security, a comprehensive long-term capacity building strategy has been launched as part of the Nuclear Security Plans 2006–2009 and 2010–2013. The chief focus of the strategy is on nuclear security education and training. To this end, several important initiatives have been undertaken. For example, in 2010 the IAEA published a guiding document entitled Educational Program in Nuclear Security which aimed to facilitate the development of sustainable nuclear security human resource development programs (IAEA 2010a). The guide outlined a model of a Master of Science (M.Sc.) and a Certificate Program in Nuclear Security in order to provide member states with a comprehensive strategy for the implementation of nuclear security education and also to encourage universities and other educational institutions to develop relevant academic programs. Since then the number of universities offering accredited courses and degrees in Nuclear Security has been steadily increasing.

In March 2010, the IAEA actively supported the establishment of the International Nuclear Security Education Network (INSEN), as the Network was

considered an important and suitable mechanism to support and promote the sustainable implementation of nuclear security education. By design, the INSEN is a partnership between the IAEA and educational and research institutions and competent authorities. Its mission is to “enhance global nuclear security by developing, sharing and promoting excellence in nuclear security education” (IAEA 2010b). The Network’s activity covers a set of key areas, including but not limited to development of peer-reviewed textbooks, computer-based teaching tools, and instructional materials; faculty development through exchanges and training; joint research and sharing of scientific knowledge and infrastructure; and quality evaluation and assurance (IAEA 2010b). So far the INSEN comprises educational and research institutions from 26 countries on 4 continents, as well as 2 international organizations. Overall, the nuclear security education experience constitutes a valuable model for the development and implementation of biosecurity education for the life sciences, insofar as it is state led, all inclusive, and internationally coordinated based on a long-term strategic planning and clear targets.

In order to prevent the large-scale militarization of the life sciences, the responsibilities of researchers should be seen within the broader framework of the “web of preventive policies” centered on the BTWC. It is essential that practicing life scientists at all levels become fully aware of their responsibilities under the Convention, so that they can contribute their expertise to strengthening the biological weapons nonproliferation regime. To this end, the implementation of a comprehensive biosecurity *education* plan at international level based on the model of the nuclear security education experience could be viewed as a vital step in the process of fostering biosecurity *norms*. Only in this way will it be possible to ensure that attempts at the hostile misuse of life science knowledge and materials are effectively discouraged and prevented and that the life sciences continue to generate benefits for peaceful, prophylactic, and preventive purposes.

Cross-References

► Biosecurity Education and Awareness in Neuroscience
► International Legal Restraints on Chemical and Biological Weapons
► Neuroscience Advances and Future Warfare

References

Australia, Canada, Japan, New Zealand, Republic of Korea and Switzerland (on behalf of the “JACKSNNZ”), Kenya, Pakistan, Sweden Ukraine, the United Kingdom of Great Britain and Northern Ireland and the United States of America. (2011). Possible approaches to education and awareness-raising among life scientists. BWC/CONF.VII/WP20/Rev.1. United Nations. http://daccess-dds-ny.un.org/doc/UNDOC/GEN/G11/643/57/PDF/G1164357.pdf?OpenElement. Accessed 7 Dec 2012.

Becker, G. S. (1996). *Accounting for tastes*. Cambridge, MA: Harvard University Press.

Butler, D. (2012). Mutant-flu researcher plans to publish even without permission. *Nature News*. http://www.nature.com/news/mutant-flu-researcher-plans-to-publish-even-without-permission-1.10469. Accessed 12 July 2013.

Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction (BTWC). (1972). Full text of the Convention available at http://www.opbw.org/. Accessed 18 Dec 2012.

Dando, M. (2010). How the 7th Review Conference of the 2011 BWC can improve life scientists' understanding of biosecurity and the dual use dilemma. Seminar delivered as part of the *Science and global security series,* University of Princeton. http://www.princeton.edu/sgs/seminars/biosecurity/archives/. Accessed 12 July 2013.

Dando, M., & Rappert, B. (2005). *Codes of conduct for life sciences: Some insights from UK Academia*. Briefing paper no.16 (2nd series). University Bradford. www.brad.ac.uk/acad/sbtwc. Accessed 1 Nov 2010.

Editorial. (2012a). For better or worse. *Nature, 484*, 415.

Editorial. (2012b). Publishing risky research. *Nature, 485*, 5.

Enserink, M., & Malakoff, D. (2003). The trials of Thomas Butler. *Science, 302*(5653), 2054–2063.

Federation of American Scientists. (2006). In support of Butler, T. C. Resource documents. http://www.fas.org/butler/index.html. Accessed 18 Dec 2012.

International Atomic Energy Agency. (2008). Nuclear security culture: Implementing guide. IAEA nuclear security series, no.7. http://www-pub.iaea.org/MTCD/publications/PDF/Pub1347_web.pdf. Accessed 7 Dec 2012.

International Atomic Energy Agency. (2010a). Educational programme in nuclear security. IAEA nuclear security series, no.12. http://www-pub.iaea.org/MTCD/publications/PDF/Pub1439_web.pdf. Accessed 7 Dec 2012.

International Atomic Energy Agency (2010b). International Nuclear Security Education Network (INSEN). http://www-ns.iaea.org/security/workshops/insen-wshop.asp. Accessed 7 Dec 2012.

National Research Council. (2004). *Biotechnology research in an age of terrorism*. Washington, DC: National Academies Press.

National Research Council. (2006). *Globalization, biosecurity, and the future of the life sciences*. Washington, DC: National Academies Press.

Novossiolova, T., & Whitby, S. (2011). Building capacity in dual-use bioethics: Biosecurity education for life scientists. *New Security Learning,* 2. http://www.newsecuritylearning.com/index.php/feature/78-building-capacity-in-dual-use-bioethics-biosecurity-education-for-life-scientists-. Accessed 13 July 2013.

Perez, D. (2012). H5N1 debates: Hung up on the wrong questions. *Science, 335*(6070), 799–801.

Posner, R., & Rasmusen, E. (1999). Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics, 19*, 369–382.

Rappert, B. (2007). *Biotechnology, security and the search for limits: An inquiry into research and methods*. Basingstoke: Palgrave.

Rappert, B. (2009). *Experimental secrets: International security, codes, and the future of research*. Lanham: University Press of America.

Tamanaha, B. (2001). *A general jurisprudence of law and society*. Oxford: Oxford University Press.

The Royal Society. (2012). Brain waves module 3: Neuroscience, conflict and security. Report. The Royal Society. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/brain-waves/2012-02-06-BW3.pdf. Accessed 20 Dec 2012.

United Nations Security Council. (2004). UNSC Resolution 1540. http://www.un.org/en/sc/1540/. Accessed 18 Dec 2012.

Vickers, G. (1972). The management of conflict. *Futures, 4*(2), 126–141.

Vickers, G. (1973). Values, norms and policies. *Policy Studies, 4*, 103–111.

Whitby, S. (2012). Strengthening the biological weapons convention: Preserving academic and scientific freedom. *Science, People and Politics, 3*(2). http://www.gavaghancommunications.com/sppwhitby.html. Accessed 7 Dec 2012.
Whitman, J. (2009). *The fundamentals of global governance*. Basingstoke: Palgrave.
World Health Organisation. (2005). International health regulations (2nd ed.). http://whqlibdoc.who.int/publications/2008/9789241580410_eng.pdf. Accessed 18 Dec 2012.

Neuroethics of Warfare

117

Gerald Walther

Contents

Abstract

International Humanitarian Law was established in the twentieth century to reduce the impact of warfare on civilians as well as to increase the protection of combatants. This chapter critiques its specific focus on the reduction of physical harm, while it does not take into consideration the impact of psychological harm onto civilians and combatants. Advances in neuroscience have shown that psychological well-being is linked with neuroanatomy and thus harm to the psyche is comparable to and even congruent with neurophysiological damage. The effects of stress and fear, which are abundant in modern warfare, are felt by combatants and noncombatants alike and have detrimental long-term effects. In addition, modern military doctrines, such as "Shock and Awe," used in the Iraqi invasion in 2003, specifically try to inflict fear and anxiety. While it may be argued that this leads to fewer casualties, the long-term effects of this doctrine may have unforeseen future consequences as societies suffer from

G. Walther
Division of Peace Studies, University of Bradford, Bradford, UK
e-mail: G.Walther@Bradford.ac.uk

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4_143,

neurophysiological damage. Neuroethics is uniquely suited to critique current military doctrines and technologies and push for a revision of International Humanitarian Law to take into consideration and apply the insights about the human mind that neuroscience has uncovered.

Introduction

After the preceding chapters in this section have dealt with future interest in neuroscience research by the military, the legal implications of the use of these technologies, the education and awareness among scientists, and how norms can be developed in epistemic communicates such as neuroscience, this final chapter, instead of simply rounding off the section, takes the rather presumptuous title of "The Missing Neuroethics." What ethics is still missing? What has not yet been covered in this section?

The central topic of discussion within the chapter will be the relationship between fear, stress, pathologies, and military warfare. Starting point is a short summary of International Humanitarian Law, specifically *ius in bello*, and how it attempts to protect both combatants and civilians in warfare. Out of this overview, the argument is made that this protection is very limited and does not take into consideration the advances that neuroscience has made in understanding the relationship between psychological events, such as fear and stress, and its impact on human health, specifically neuropathologies. However, modern military doctrines such as "Shock and Awe," as used in the Iraqi invasion in 2003, as well as military interest in neuropharmacology, might increase the likelihood that both combatants and civilians become increasingly subjected to neuropathologies. Finally, the paper argues that neuroethics is uniquely suited to inform and change International Humanitarian Law to include psychological harm by showing its relationship with neuroanatomy. It can also inform neuroscientists how their research could be used by the military and the impact it can have on human security.

Contention: A Void in Humanitarian Law

As long as humans have engaged in warfare, there have always been calls for restraint on the methods of warfare. Within the Western world, the discourse on ethics and warfare was initiated by Christian theology. Early Christianity adopted a pacifist approach; however, as Johnson has argued, this ideology became more and more difficult to follow as time progressed and the Second Coming of Christ did not occur (Johnson 1987). As Christianity spread and soldiers within the Roman army converted, it became imperative to develop a Christian understanding of war that goes beyond pacifism and determines when a Christian is allowed to take up arms. In response to this need, St. Augustine of Hippo (354–430) developed the first theory of just war and started to differentiate between *ius ad bellum* and *ius in bello*,

i.e. just cause for war and justice in war's conduct, respectively. The issue was again taken up in the thirteenth century by St. Thomas Aquinas in his exploration of the justification of the Crusades. Eventually, Grotius (1583–1645) secularized the theory in his book on "The Laws of War and Peace," which is an analysis of international law as informed by natural law. As modern states started to emerge in the seventeenth century and wars were waged on national rather than local levels, the idea of an international law governing states received more attention. In the mid-nineteenth century, states started to develop formalized agreements on *ius in bello* issues, which are of primary relevance to this chapter, starting with regulations on how to treat wounded and sick soldiers in the field, which was later extended to armed forces at sea as well. The last addition after WWI pertained to the treatment of prisoners of war. After WWII all three agreements were reevaluated, and the protection of civilians in times of war was included. These agreements are the Geneva Conventions, which have since then been amended by additional three protocols: Protection of Victims of International Armed Conflicts, Protection of Victims of Non-International Armed Conflicts, and the Adoption of an Additional Distinctive Emblem. So what are some of these principles that are to be followed when engaging in warfare? How and to what degree does humanitarian law protect combatants and noncombatants?

In order to fight a war justly, the following principles need to be observed (Evans 2005):

1. Discrimination: One should not directly target those who do not directly participate in the war, and, in addition, one should take precaution to limit the casualties among these nonparticipants.
   (a) Doctrine of double effect: However, if an action that is justified by other criteria of the just war theory entails the deaths of nonparticipants, i.e., they are unintentional, then the war is not unjust.
2. Proportionality: The use of force deployed to achieve an objective has to be proportionate.
3. All noncombatants, which include prisoners of war, have to be treated justly.
4. All other international and national laws need to be followed as long as they are not fundamentally in conflict with the theory's moral requirements.

Of primary concern for this chapter is the first point of discrimination. It is more clearly defined in the Protocol Additional to the Geneva Convention relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, specifically in Part IV, Civilian Population. Articles 48 through 51 outline who is considered a civilian and declare that civilians are granted special protection within war, e.g. that they cannot be objects of military actions, that indiscriminate attacks are prohibited, that they cannot be specifically targeted in war, and that any acts to spread fear among them are not allowed. In general, the goal of these provisions is to ensure that "[t]he civilian population and individual civilians shall enjoy general protection against dangers arising from military operations" (Art. 51.1). But what exactly comprises dangers? Evans' analysis of just war theory talks about avoidance of casualties with regard to the issue of discrimination in war (Evans 2005).

In his seminal work, "Just and Unjust Wars," Walzer uses historical examples to highlight certain issues within the just war theory and International Humanitarian Law (Walzer 1977). All of his examples related to civilian protection focus on physical harm, i.e. mostly people dying due to decisions of military commanders and the analysis discusses if these decisions were in accordance with humanitarian law. In his historical overview of humanitarian law, Solf claims that one of the four elements inherent in humanitarian law is that "loss of life and destruction of property must have some rational tendency to the prompt achievement of a definite military advantage" (Solf 1986). What is lacking in this evaluation of the "justness" of warfare is the psychological component and the damage that results from various types of military technologies and strategies. Of course, at the time of writing of these rules to protect civilians as well as combatants, our understanding of psychological pathologies had been limited. But this is where neuroethics will be able to make a fundamental contributions to the literature on human rights and international law by critiquing the emphasis on the purely physical dimensions of harm. As we progress in our understanding of the relationship between fear, anxiety, trauma, and pathologies, we may need to revise our definition of what harm is and what can be considered "acceptable" harm. And this analysis, although preeminent to the military warfare, stretches across other domains as well, for example, exercising control in prisons. In order to demonstrate where this type of analysis can lead us, this chapter will address the relationship between fear, anxiety, and trauma. Fear is particularly interesting because it is a tool that has been part of the military arsenal from traditional warrior societies to the modern military complex. And with our advances in understanding fear, the military will become increasingly capable to deploy it in even more sophisticated ways, e.g. using psychopharmacological agents. The next part of this chapter will look into our current understanding of the relationship between fear, anxiety, trauma, stress, and pathologies.

The Neuroscience of Fear and Its Psychophysiological Effects

What Is Fear?

While in our everyday life, the use of the word "fear" is quite frequent, and people appear to understand what is meant by it, which is to express some sort of experience of apprehension, in the scientific world, "fear" is rather ill defined and murky. As Rachman points out, it is very difficult to get people to talk about fears because of social stigma (Rachman 1978). For example, in warfare, soldiers are discouraged to think and admit their fears. And even absent a significant event like war, men are quite often culturally expected not to show signs of fear. Sometimes patients who suffer from fear are even unable to identify and recognize that they are afraid. These issues of admittance or recognition impair scientific study. But the study of fear also encounters a semantic barrier. People occasionally say that they are afraid of a certain event or object. However, once they are faced with this object,

they may not express any symptoms of fear at all. According to Rachman, it is also very difficult to quantify fear, because how does one translate expressions such as "terrified," "anxious," or "very frightened." As a result, a lot of work on fear has focused on the physiological and observable behavioral changes that occur when people say that they are afraid or in fear. Following Rachman, it is thus more useful to describe fear in terms of three components: "the subjective experience of apprehension, associated psychophysiological changes, and attempts to avoid or escape from fearful situations" (p. 3). However, as Rachman continues, these "three components often fail to correspond" (p. 3), which makes it "helpful to specify which component of a fear one is describing" (p. 3). Another distinction when talking about fear is between acute and chronic types. An acute fear might be, for example, triggered by seeing a snake, whereas chronic fears can either have a direct stimulus, e.g. the fear of being alone, or not. These latter are sometimes labeled anxieties that are more of an apprehension (Rachman 1978). Anxiety comes from within us (LeDoux 1999). Clinically speaking, fear and its accompanying physiological changes are adaptive in that they help us prepare for a specific danger, whereas anxiety is maladaptive and prevents us from functioning properly. Therefore, there are only anxiety disorders, but never fear disorder. According to the Diagnostic and Statistical Manual of Mental Disorder, Fourth Edition, Text Revision (DSM-IV-TR), the following anxiety disorders are classified: generalized anxiety disorder (GAD), social anxiety disorder (also known as social phobia), specific phobia, panic disorder with and without agoraphobia, obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), anxiety secondary to medical condition, acute stress disorder (ASD), and substance-induced anxiety disorder (American Psychiatric Association 2000). According to Nutt et al., while these are specific disorders, each exhibits "shared features such as anxiety, certain physiological features and behavioural changes... These central features are also present in and indeed form the core of the other anxiety disorder-specific phobia," which is the equivalent of fear responses (Nutt et al. 2008, p. 365).

Given the varieties of fear and anxiety dimensions, it seems best to heed Rachman's advice to first describe what specific concept or version of fear one is investigating. This chapter will look at the psychophysiological dimension of fear and anxiety. However, instead of trying to understand what fear is, and thus differentiate between fears and anxieties, the focus is on elucidating how fears and anxieties manifest themselves and what their long-term effects are on the individual. In Rachman's description of the psychophysiological variables, i.e., "perspiring, trembling, or increased heart rate," that are employed to measure fear, it is noticeable that his book is slightly outdated – it was first published in 1978. Today, advances in cognitive neuroscience allow for a more detailed understanding of the psychophysiological changes associated with fear in terms of brain chemistry. It also allows for an inclusion of the concept of trauma and its relationship with fear, anxiety, and brain pathologies, which will be discussed in the last part of this section. The next part addresses the questions of what fear, anxiety, and trauma are in terms of neurophysiology.

The Neurophysiology of Fear and Anxiety

What fear and anxiety certainly share in common is the component of stress. When I find myself in a position of being afraid, it puts stress on my psychophysiological system. Given the larger context of this paper on the psychological effects of warfare, military technologies, and control, it is also necessary to look at the role of stress. So what is stress and what does it do?

Stress is brought on by aversive or dangerous situations, in which our system prepares our body to deal with the effects of a fight-or-flight scenario, which is either to fight the adversary or to run away. Normally, the accompanying changes to our system are only brief because running away and fighting are very short events. However, if there is no behavioral solution to a fight-or-flight event, we find ourselves in a state of prolonged stress response. This prolonged state can fundamentally affect our physiology and lead to pathologies that last for months or years. Carlson explained the neurophysiological effects of stress in his widely used textbook on "Physiology of Behavior," (Carlson 2013) and the following sections synthesize his discussion of this topic.

All emotions generally consist of a behavioral, autonomic, and endocrine response. Of particular interest to this chapter are the autonomic and endocrine responses. Of course, behavioral reactions to stress can also be dangerous, e.g., if a soldier under artillery fire runs out of the trenches because he wants to get away from the explosions. However, these are individual instances that are detrimental to the individual who displays such "inappropriate" behavior. But the autonomic and endocrine responses are shared by everyone in a stressful situation and are thus more important for this chapter. Both autonomic and endocrine responses are catabolic because they enable the body to mobilize all of its resources. The body activates the sympathetic branch of the autonomous nervous system, and the adrenal glands secrete epinephrine, norepinephrine, and steroid stress hormones. The results of the release of these hormones are manifold: epinephrine causes the nutrients in our muscles to be released in order to provide additional energy; it also, together with norepinephrine, increases the blood flow to the muscles and thus the overall blood pressure; cortisol, a steroid and glucocorticoid, helps to break down proteins and convert them into glucose, which allows fats to be made available for energy, increases blood flow, and affects the behavior by interacting with the brain. It also reduces the secretion of sex steroid hormones. As nearly all cells in the body contain glucocorticoid receptors, its effects are far from being understood completely. The release of glucocorticoids is mediated by neurons in the periventricular nucleus of the hypothalamus, which secrete a peptide called corticotropin-releasing hormone (CRH), which in turn stimulates the anterior pituitary gland, which in turn releases another hormone that eventually stimulates the adrenal gland. CRH is also secreted in the brain and has an effect on the limbic system, which is largely responsible for emotional responses and contains the periaqueductal gray matter, the locus coeruleus, and the central nucleus of the amygdala. Studies have shown that injection of CRH into the brain increases the acquisition of fear-conditioned responses and anxiety in rats.

But what about the health effects of stress, particularly long-term stress, on health? Carlson describes both the physiological and the neurophysiological effects. As already mentioned, glucocorticoid receptors are present in nearly all human cells. The effects of long-term release of the hormone are thus quite varied: increased blood pressure, damage to muscle tissue, steroid diabetes, infertility, inhibition of growth, inhibition of the inflammatory responses, and suppression of the immune systems. In addition to these primary effects, medical solutions to some of the issues cause secondary problems. For example, the long-term use of steroids to treat inflammations can result in cognitive deficiencies and lead to steroid psychosis, which includes several symptoms: profound distractibility, anxiety, insomnia, depression, hallucinations, and delusions. While all of these are very detrimental to the human being, it is even more important to understand what recent neuroscientific research has discovered about the neurophysiological effects of stress. Research has particularly focused on a few specific issues: glucocorticoid and neuron interaction, prenatal stress, early-life stress, and brief exposure to stress, all of which are coupled with high-stress and low-stress stimuli:

- General, long term: In animals, long-term exposure to glucocorticoids destroys neurons in the CA1 of the hippocampal formation. As this region is involved in memory and learning, stress might have permanent implications for these two cognitive abilities. Among elderly people, it has been shown that those with high blood levels of glucocorticoids learned a maze slower than those with normal levels.
- Prenatal, long term: In rats, prenatal stress interferes with hippocampal development as well, which impedes learning, memory, and spatial recognition tasks. In addition to its impact on the hippocampus, even mild prenatal stress produced permanent changes to the amygdala of unborn rats, which leads to a heightened sensitivity to fear.
- General, short term: Placing a rat in a plexiglass box into a cage with a cat for 75 min resulted in permanent changes in their neurophysiology. The glucocorticoid level increased five times, which caused changes in their spatial learning ability, and their primed-burst potentiation, which is a form of long-term potentiation, was reduced in their hippocampal slices. In addition, acute stress has also been shown to reduce the long-term survival of hippocampal neurons that are produced in the process of neurogenesis, which, when it is impaired in the hippocampus, has been linked with depression.
- Primates: While a lot of these studies have been conducted using rats, it has been shown that among vervet monkeys, who have a hierarchical social order, in which the ones at the bottom are continuously subjected to stress, some of these monkeys have died of stress. They suffered from gastric ulcers and increased adrenal glands, and in terms of their neuroanatomy, the neurons in their CA1 field of the hippocampal formation had been completely destroyed. Among humans who have been subjected to torture, brain degeneration has been documented with CT scans. In humans, even mild stress in early life reduced the volume of the dorsomedial prefrontal cortex. Similarly, chronic pain patients have reduced gray

matter in their cerebral cortex and particularly in the dorsolateral prefrontal cortex. In cognitive terms, chronic pain patients perform similarly poor on a task as those who have been affected by prefrontal lesions.

Post-traumatic stress disorder (PTSD) has also received widespread attention in the scientific community. According to the DSM-IV-TR, PTSD is caused when a person "experienced, witnessed, or was confronted with an event or events that involved actual or threatened death or serious injury, or a threat to the physical integrity of self or others" and "the person's response involved intense fear, helplessness, or horror". Its psychological symptoms include difficulty falling asleep or staying asleep, irritability, outbursts of anger, difficulty in concentrating, and heightened reactions to sudden noises and movements. In addition to these mental health implications, it also leads to poor physical health (Zayfert et al. 2002). In terms of the neurophysiological effects, they are very similar to those of a high-stress environment. Thus, specifically the hippocampus as well as the amygdala is affected. It has also been shown that the probability of suffering from PTSD positively correlates with the number of traumatic events that have been witnessed. This finding is supported by the evidence found in analyzing the hippocampal region of monozygotic twins, of whom one has fought in the Vietnam War. The war-exposed twins with PTSD had smaller hippocampal regions than those war-exposed twins without PTSD. Remarkably though, the PTSD soldier's twin who had not been in combat also had smaller than average hippocampal regions. Thus, the lower hippocampal volume is a result of the traumatic events, but people with a smaller initial hippocampal region are also more likely to develop PTSD. And if someone is exposed to traumatic events repeatedly, the hippocampal region decreases with each event and eventually reaches the threshold when symptoms start to appear.

The initial focus of this section was on fear and anxiety, which was then turned into a discussion about the effects of stress. Of course, stress in the animal research that has been discussed has been induced in the animals by creating a situation where the animal has been put in a fearful situation, e.g. exposing a rat to a cat. While the result of the physiological effects of producing a fight-or-flight scenario has been helpful for our survival, this last section has shown that even these short bursts have a negative impact on our neurophysiology. However, the most problematic changes only occur in prolonged states of stress or high-stress traumatic events. Going back to the earlier distinction between anxiety and fear, the former being considered maladaptive and the latter useful, it is scientifically warranted to lose this distinction (Blanchard and Blanchard 2008). Blanchard and Blanchard argue that there is no biological basis to support this normative judgment. Both animal and human defense reactions (Blanchard et al. 2001) to an adverse scenario are tightly controlled by the specific features of the danger and very consistent and robust (Blanchard and Blanchard 2008). Thus, anxiety and fear behavior may just be the two strategies that have enabled us to best survive in dangerous situations. For example, in rats, the direct exposure to a cat elicits a fear response, whereas putting cat hair or odor next to the rat results in the display of an anxious behavior (Adamec et al. 1998). Both may be useful in the

short run, i.e. to survive, but detrimental in the long run, i.e. the effects of stress on the physiology and neurophysiology as discussed previously.

The Hidden Harm of Warfare and Violence

In the 2003 invasion of Iraq, the US military used a strategy called "Shock and Awe," which was developed for the National Defense University by the Defense Group Inc. already in 1996 (Ullman and Wayde Jr 1996). The goal of "Shock and Awe" is not to use overwhelming force to defeat the enemy, the primary doctrine of the USA at that time, but rather to undermine the willingness of the enemy soldiers to continue or even start to fight. In the Iraqi invasion, this was achieved by hitting Baghdad with several hundred cruise missiles to specifically target command headquarters instead of actual weapons systems. As one pentagon official remarked prior to the invasion, "there will be not be a safe place in Baghdad" (CBS News 2009). Of course, from the point of view of the invading army, in this case the USA, it is certainly advantageous to reduce the morale of enemy soldiers and maybe not have them fight at all. It may even be argued that it leads to fewer casualties. However, as we have seen from the evidence on how stress results in neurophysiological damage, it is not the case that there is no harm done using this method. In fact, this warfare does not differentiate between soldiers and civilians at all. "Shock and Awe" targets and affects everyone. Of course, the effect will be different on the specific populations, e.g., soldiers, civilians, and children. As we have seen in the studies on prenatal and early-life stress on rats, children are especially vulnerable as the impact of stress directly impedes their natural neurodevelopment and cause lifelong mental deficiencies. Research has shown that negative events in early life increase the chance of psychopathologies such as post-traumatic stress disorder (Paolucci et al. 2001), social anxiety (Binelli et al. 2012), or general anxiety disorder (Phillips et al. 2005). The effect on soldiers might also be heightened compared to a normal civilian for a variety of reasons. Warfare is highly ritualized. Van der Dennen's discussion and analysis of warfare among traditional societies show the various types of rituals that are conducted before as well as after the actual combat (Van der Dennen). Most of the pre-warfare rituals serve to protect the warriors against harm, e.g. praying for the warriors, giving the warriors charms and amulets, divination, asceticism, sacrifice, vows of the warriors to be victorious before the community, decorations like body painting, or rehearsal of the battle, which are enacted in advance and show that they will be victorious. However, these rituals do not only try to persuade and convince the warrior that he will not be harmed. They serve as a reminder of the duty to protect their society, as a means to enhance the belief in one's own strength and to protect against fear and instill fear in the opponent. While soldiers hardly face each other directly in modern warfare anymore, instilling fear in the enemy is achieved via superior military technology or strategy, e.g., the "Shock and Awe" doctrine. In addition, what distinguishes the modern soldier from a traditional warrior is that the warrior retains the right to enter into negotiation and talk to the enemy. To a modern soldier, this possibility does not

exist. Thus, the only solution to escape the fear and stress of combat is to flee, which is desertion for the individual soldier. Of course, since the soldier has undergone the blessings of his community, this action carries a high social stigma. He did not only abandon his fellow soldiers but also the community for whom he was fighting. Thus, in addition to potential neurological damage due to stress, the soldier also faces additional stress due to his failure as a soldiers in the eyes of his comrades and the community, which could result in further damage.

A military doctrine like "Shock and Awe" is only one aspect of modern warfare and violence that societies face. Post-traumatic stress disorder does not only happen to the soldiers of the losing military. The victorious military and its society also face the burden of having to deal with these traumatized, and as we have seen, actually brain-damaged soldiers. Also, terrorism is in its psychological effects similar to the "Shock and Awe" theory in that it tries to show those societies that the terrorists are "in war" with that no one in the society is safe from their attacks. Thus, if a suicide bomber blows himself up, there are three categories of victims: those directly harmed or killed in the attack, the bystanders who witnessed the attack, and the general society who have to be afraid that they could be a potential victim next time. Members in the two last groups could equally suffer from neurological damage and accompanying behavioral changes. While military warfare and terrorism are more prone to cause the types of stress and damage that have been discussed, the issue does not only pertain to those two. For example, victims of torture show brain degeneration in CT scans (Carlson 2013). Similarly, solitary confinement of prisoners may also cause behavioral and neurophysiological changes. For example, in rabbits it has been shown that rabbits raised individually in a cage show higher fear levels and incomplete behavioral patterns than their peers who were raised in a group (Trocino et al. 2012).

Conclusion

International Humanitarian Law has been developed in the last century to protect civilians and soldiers in warfare. However, its focus on physical safety may need to be enlarged to account for the neurophysiological damage that can be caused by stress, which is a product of fear and anxiety. These neurophysiological changes can be permanent and may be highly detrimental for society to return to or transition into a peaceful society. Modern military doctrines like "Shock and Awe," which was applied in the 2003 Iraqi invasion, specifically try to create fear in the enemy soldiers, which nevertheless also target the civilian population to an equal degree. As neuroscience advances, it is likely that military technology will be better able to produce fear in the enemy, for example, via neuropharmacological agents. For example, while fear inducement is not specifically mentioned in a 2008 report by the "Committee on Military and Intelligence Methodology for Emergent Neurophysiological and Cognitive/Neural Science Research in the Next Two Decades," the use of technologies for degradation, i.e. the decrease of the abilities and skills of the enemy, is considered a field of potential benefit for the military and the intelligence community (National Research Council 2008). Neuroethics is

uniquely positioned to inform the international community of the danger of fear inducement and to show the hidden costs of war and violence. It can also raise awareness among neuroscientists of the need to be informed about the possibility of misuse of their research and how it can impact on military warfare.

Cross-References

▶ International Legal Restraints on Chemical and Biological Weapons
▶ Neuroscience Advances and Future Warfare
▶ Weaponization of Neuroscience

References

Adamec, R., Kent, P., Anisman, H., Shallow, T., & Merali, Z. (1998). Neural plasticity, neuropeptides and anxiety in animals – Implications for understanding and treating affective disorder following traumatic stress in humans. *Neuroscience & Biobehavioral Review, 23*, 301–318.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*. (4th ed, text rev.) Washington, DC: American Psychiatric Association.

Binelli, C., Ortiz, A., Muñiz, A., Gelabert, E., Ferraz, L., Filho, A. S., Crippa, J. A. S., Nardi, A. E., Subirà, S., & Martín-Santos, R. (2012). Social anxiety and negative early life events in university students. *Revista Brasileira de Psiquiatria, 34*(Supl. 1), 569–580.

Blanchard, D. C., & Blanchard, R. J. (2008). Defensive behaviors, fear, and anxiety. In R. J. Blanchard, D. C. Blanchard, G. Griebel, & D. Nutt (Eds.), *Handbook of anxiety and fear* (pp. 63–79). Amsterdam: Elsevier.

Blanchard, D. C., Hynd, A. L., Minke, K. A., Minemoto, T., & Blanchard, R. J. (2001). Human defensive behaviors to threat scenarios show parallels to fear- and anxiety-related defense patterns of non-human mammals. *Neuroscience and Biobehavioral Reviews, 25*, 761–770.

Carlson, N. R. (2013). *Physiology of behavior* (11th ed.). Princeton: Pearson Education.

CBS News. (2009). http://www.cbsnews.com/stories/2003/01/24/eveningnews/main537928.shtml (the article was originally published by CBS News on 24 Jan, 2003).

Evans, M. (2005). *Just war theory: A reappraisal*. Edinburgh: Edinburgh University Press.

Johnson, J. T. (1987). *The quest for peace: Three moral traditions in western cultural history*. Princeton: Princeton University Press.

LeDoux, J. E. (1999). The emotional brain: the mysterious underpinnings of emotional life. London: Phoenix.

National Research Council. (2008). *Emerging cognitive neurosciences and related technologies*. Washington, DC: National Academies Press.

Nutt, D., Garcia de Miguel, B., & Davies, S. J. C. (2008). Phenomenology of anxiety disorders. In R. J. Blanchard, D. C. Blanchard, G. Griebel, & D. Nutt (Eds.), *Handbook of anxiety and fear* (pp. 365–393). Elsevier: Amsterdam.

Paolucci, E. O., Genuis, M. L., & Violato, C. (2001). A meta-analysis of the published research on the effects of sexual child abuse. *Journal of Psychology, 135*(1), 17–36.

Rachman, S. J. (1978). Fear and courage (2nd ed.). New York: Freeman.

Solf, W. A. (1986). Protection of civilians against the effects of hostilities under customary international law and under protocol I. American University *Journal of International Law and Policy, 1*, 117–135.

Phillips, N. K., Hammen, C. L., Brennan, P. A., Najman, J. M., & Bor, W. (2005). Early adversity and the prospective prediction of depressive and anxiety disorders in adolescents. *Journal of Abnormal Child Psychology, 33*(1), 13–24.

Trocino, A., Majolini, D., Tazzoli, M., Filiou, E., & Xiccato, G. (2012). Housing of growing rabbits in individual, bicellular and collective cages: Fear levels and behavioural patterns. *Animals*. doi:http://dx.doi.org/10.1017/S1751731112002029.

Ullman, H., & Wayde Jr., J. P. (1996). *Shock and awe – Achieving rapid dominance*. National Defense University, Center for National Strategic Studies. http://www.dodccrp.org/files/Ullman_Shock.pdf

Van der Dennen, J. M. G. Ritualized "primitive" warfare and rituals in war: Phenocopy, homology, or...? http://rint.rechten.rug.nl/rth/dennen/ritual.htm. Accessed 10 Jan 2013.

Walzer, M. (1977). *Just and unjust wars: A moral argument with historical illustrations* (4th ed.). New York: Basic Books.

Zayfert, C., Dums, A. R., Ferguson, R. J., & Hegel, M. T. (2002). Health functioning impairments associated with posttraumatic stress disorder, anxiety disorders, and depression. *The Journal of Nervous and Mental Disease, 190*, 233–240.

Index

J. Clausen, N. Levy (eds.), *Handbook of Neuroethics*,
DOI 10.1007/978-94-007-4707-4,

E